PHYSICAL DESIGN SOLUTIONS FOR 3D ICS AND THEIR NEUROMORPHIC APPLICATIONS

A Dissertation
Presented to
The Academic Faculty

By

Bon Woong Ku

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology

May 2019

PHYSICAL DESIGN SOLUTIONS FOR 3D ICS AND THEIR NEUROMORPHIC APPLICATIONS

Approved by:

Dr. Sung Kyu Lim, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Saibal Mukhopadhyay
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Arijit Raychowdhury
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Tushar Krishna
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Hyesoon Kim
College of Computing
Georgia Institute of Technology

Date Approved: March 14, 2019

ACKNOWLEDGEMENTS

While successfully making it to the finish line of my Ph.D. life, I would like to say that, in retrospect, this bumpy ride has been full of joy and happiness because I have met so many great people who have inspired me all the time. I would like to take this opportunity to express my gratitude to all of them.

First, I would like to thank my advisor, Dr. Sung Kyu Lim, for his huge support and great guidance on my research. He has encouraged me to be on the right track, and boosted me up all the way so that I could take one of great achievements in my life successfully. Also, I would like to thank Dr. Saibal Mukhopadhyay and Dr. Arijit Raychowdhury for guidance and suggestions on my research. In addition, I thank Dr. Tushar Krishna and Dr. Hyesoon Kim for their time serving my dissertation defense committees.

I appricate IMEC, Synopsys, and Intel Labs: Praveen Raghavan, Peter Debacker, Arthur Nieuwoudt, Young-Joon Lee providing with technical resources, discussion, and guidance regarding my research. I thank GTCAD members: Dr. Shreepad Panth, Dr. Taigon Song, Dr. Yarui Peng, Dr. Sandeep Samal, Kyungwook Chang, Anthony Agnesina, Sai Pentapati, Jinwoo Kim, Da Eun Shim, Jee Hyun Lee, Yi-Chen Lu, Lingjun Zhu, Gauthaman Murali, Chengjia Shao, Rakesh Perumal for their valuable comments and feedback.

I appreciate all my friends in Atlanta supporting and cheering me up all the time. Without their positive influence, I believe that this long journey could be endless and worthless. So many precious moments with them would be remembered forever in my life.

Finally, my deepest gratitude goes to my parents Mrs. Yeon Ok Park and Mr. Ja Phil Ku, and my brother Bon Wook Ku. This phase of my life would not have been succesfully done without their encouragement, unconditional love, constant support, and patience.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The wafer-level 3D integration including face-to-face (F2F) and monolithic 3D (M3D) technologies has been featured as a promising innovation to succeed the horizontal device scaling benefit in the looming end of Moore's law. While through-silicon-via-based 3D integration requires a huge silicon-area overhead to make 3D connections between separate tiers, the wafer-level 3D integration enables fine-grained vertical interconnections down to the transistor-level. This allows physical designers the higher degrees of freedom in 3D placement and routing (P&R) than any other 3D integration approaches, which maximizes the power-performance-area (PPA) benefits of 3D ICs.

The objective of this research is two-fold: Firstly, to develop computer-aided-design (CAD) methodologies to address potential issues of the wafer-level 3D integration including power integrity, inter-tier variations, and cost overhead. Secondly, to evaluate the PPA benefits of the wafer-level 3D integration to the neuromorphic processor design at the full-chip level by applying proposed solutions.

For the first part, the static power integrity issue of transistor-level M3D ICs is inspected in detail, and we address the issue by proposing a new layout scheme for transistor-level M3D standard cells. Next, physical design solutions for gate-level M3D ICs are developed to mitigate the negative impact of inter-tier device and interconnect variations, as well as the cost overhead issue. In addition, we present the unique physical design solution named Compact-2D flow, which produces commercial-quality gate-level F2F IC layouts. For the second part, we adopt the liquid-state-machine architecture, a model of recurrent spiking neural networks, to build an online machine-learning hardware platform, and study the PPA benefits of gate-level F2F and M3D ICs on the non-trivial real-world speech recognition application. This work serves as an important step towards realizing bio-inspired neuromorphic processors utilizing 3D IC design advantages.

CHAPTER 1
INTRODUCTION

Since Gordon Moore made his prediction in 1965 [1], the feature size of a transistor has continuously shrunk from $50\mu m$ to $7nm$ [2], and the industry is now looking toward the $5nm$ feature size [3]. The technology scaling has increased the number of transistors in integrated circuits exponentially, and allowed us to merge more distinct functions into a single chip with less power. Nowadays, millions of transistors are integrated per mm^2, and a single chip is made up of billions of transistors to provide better system performance [4].

While following the technology scaling roadmap, the structure and materials of a transistor have gone through multiple evolutions, such as by employing strained silicon, high-κ metal gate, and up-to-date FinFET architecture [2], to make itself smaller and perform better. Moving towards the $5nm$ and $3nm$ technology era, however, would not be the same as before because the feature sizes of the next technology nodes are physically too small to fabricate reliably at a low cost. In addition, the next generation device architecture requires an unprecedented evolution to embrace the 3D nature to provide the improved electrical performance with the extremely small feature size [5, 6]. Gate-all-around FET [7, 8, 9] targeting the $5nm$ technology node vertically stacks the multiple epitaxially-grown silicon nanowires for the channel formation to suppress the short-channel impact more effectively. Therefore, the vertical alignment among nanowires and the crystallization quality of nanowires are critical for the manufacturing process. Also, reduction in the vertical spacing between nanowires is important to decrease the gate capacitance, which implies that the geometric scaling of a transistor should happen in both the horizontal and the vertical dimensions.

While previous technological breakthroughs have been made at the device level in general, what makes the evolution of today special is that the device scaling continues in com-

bination with various vertical stacking approaches at the circuit integration level. This is because the 3D device process requires reduction in new defect mechanisms and process complexity, which is expected to hamper the scaling trend significantly. Therefore, while the device architecture explores new opportunities in the vertical dimension, the approach of circuit integration itself is now forced to change in the way of exploiting the vertical dimension.

Stacking multiple dies in a 3D fashion has also evolved in many different ways including stacking of either packaged dies (package-on-package, PoP), or bare dies (stacked-integrated-circuits, SiC) [10, 11]. However, there are still huge opportunities in the 3D integration in the context of the interconnect scaling. As 3D integration technology has matured, both the length and the pitch of 3D interconnections has become shorter. In the packaging-level 3D integration, 3D interconnects have been made by ball-grid-array and wire-bonding technologies, which is $100\mu m$-scale. In the die-level 3D integration, microbump-array and through-silicon-via (TSV) technologies, which is $10\mu m$-scale, have been used for 3D interconnects. While a $2\mu m$ pitch has been demonstrated in the advanced TSV technology, a $40\mu m$ pitch of microbump array has been the main bottleneck. Also, TSV-based 3D integration requires a die alignment step after dies are fabricated in parallel, so the size of a TSV should have its μm-scale lower bound to avoid unexpected disconnection during the die alignment process.

1.1 Wafer-level 3D Integration

Over the last few years, the wafer-level 3D integration has emerged as a promising solution to enable bumpless sub-μm 3D interconnects. Depending on how electrical connections are fabricated between the tiers, the wafer-level 3D integration is categorized into face-to-face (F2F) and face-to-back (F2B) integration. Lately, the hybrid wafer-to-wafer (W2W) bonding technology [12, 13] and the monolithic 3D technology [14, 15] have emerged as promising solutions for the advanced F2F and F2B wafer-level 3D integrations, respec-

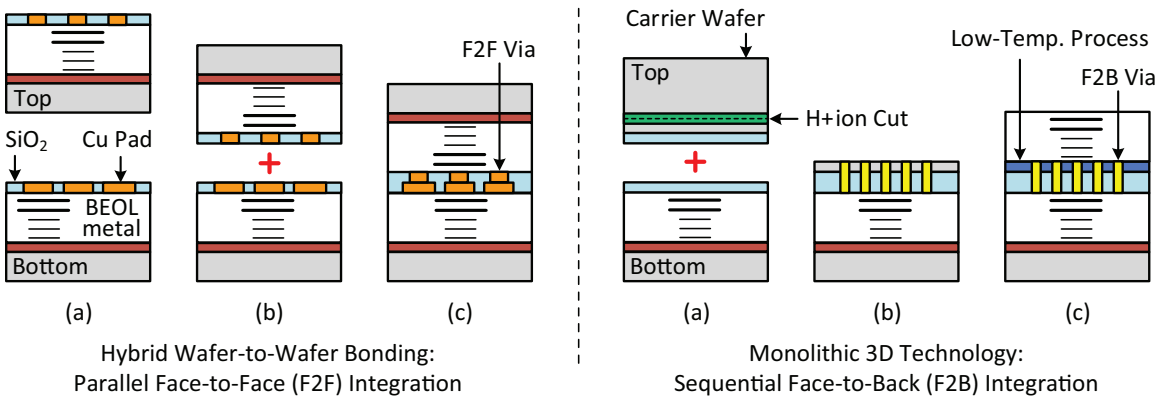


Figure 1.1: Two types of the wafer-level 3D integration technologies.

tively.

The F2F integration is enabled by the hybrid W2W bonding, which allows direct metal-to-metal (damascene-pad) and dielectric-to-dielectric bonding between the back-end-of-lines (BEOLs) of pre-fabricated wafers. Wafers are fabricated in parallel with the conventional process before bonding, and wafer surface is flattened by chemical-mechanical planarization. Then, wafers are overlapped in a F2F fashion, aligned, bonded at room temperature, and annealed at (250°C) to strengthen the inter-facial bonding. As a result, F2F vias are naturally formed at the locations of direct metal-to-metal bonding. Recently, a $1.8\mu m$ F2F via pitch has been demonstrated, and the minimum pitch is projected down to $0.8\mu m$ in the near future [16, 17].

On the other hand, the F2B integration is enabled by the monolithic 3D (M3D) technology, which allows the sequential fabrication process on top of the pre-fabricated bottom wafers. The top empty substrate is sheared off from the bulk carrier by the H+ ion implant cut, and bonded at a low temperature to the top of the bottom wafer. After planarization, F2B vias, or monolithic inter-tier vias (MIVs) are created for the 3D interconnections with the litho-scale precision, and top devices and metal layers are fabricated within a low thermal budget to keep the integrity of bottom FEOLs and BEOLs. Because of this inter-tier process variation and following fabrication cost issues, the F2B integration is farther to the

commercialization than the F2F integration, but the minimum pitch of MIVs is projected down to less than $0.1\mu m$ because the F2B integration removes the wafer alignment process.

Compared to the traditional through-silicon-via (TSV)-based 3D integration where the minimum pitch of 3D interconnections is larger than $10\mu m$, the wafer-level 3D integration technology offers higher degrees of freedom in 3D placement and routing (P&R) to designers. While F2F-bonded 3D ICs utilize the 3D interconnections up to the sub-block, or gate-level, M3D ICs enable up to the sub-gate, or transistor-level 3D interconnections with a *nm*-scale 3D interconnection pitch. As a result, both F2F and M3D integration technologies help reduce the wirelength and buffer count for the full-chip design significantly, and possibly maximize power-performance benefits of 3D ICs even more than those from logic scaling [18].

1.2 Bio-inspired Neuromorphic Computing Paradigm

Another trend we need to focus in addition to the integration-level breakthrough is that computing architecture starts to adopt the working priciples of a biological nervous system these days. Today's Internet-of-Things (IoT) paradigm handles formidable data provided from us, and demands more efficient information processing at the edge computing level. As small form factor and low power consumption for the on-chip hardware machine learning is found critical, bio-inspired neuromorphic computing has emerged to overcome the Von Neumann bottleneck and the physical limitation of technology scaling. Recently, deep learning algorithms [19, 20] have delivered the state-of-the-art performance for a wide range of applications, such as image classification and natural language processing [21, 22]. However, these networks are power-hungry and necessitate a huge amount of data and computing resources for the learning process. This makes them far from being deployed for portable computing devices.

It is widely accepted that the ultimate brain-inspired computing system would closely resemble the brain behaviors rather than replicate the high-level architectural characteristics

of the cortical circuits. Towards this goal, significant research efforts have been made on spiking neural networks (SNNs) due to their biological plausibility and energy-efficiency. Among noticeable approaches using SNNs, the liquid state machine (LSM) has been featured as a special competent for spatiotemporal pattern classification, such as speech recognition and bio-signal processing [23, 24, 25]. The LSM architecture consists of a reservoir and a readout stage. In the reservoir, neurons are randomly and recurrently connected like a liquid pond. When the input patterns are fed into the reservoir stage, interactive spike signaling between recurrently connected reservoir neurons transforms the input patterns into the high-dimensional non-linear space. These preprocessed input patterns are used for the supervised spike-timing-dependent plasticity learning in the readout stage for the final classification. Thanks to the architectural simplicity and computational efficiency, LSM processors are expected to offer a promising solution for the small form factor and low power portable neuromorphic computing devices in the IoT era.

1.3 Scope of This Dissertation

In this research, we explore various design and CAD solutions for the wafer-level 3D integration technology and apply them to build 3D neuromorphic processors. We first present an optimized transistor-level M3D (T-M3D) standard cell layout scheme named stitching scheme, and show that our layout optimization improves the full-chip power integrity as well as power-performance-area savings of T-M3D ICs. Next, we study the cost and inter-tier variation impacts on gate-level M3D (G-M3D) ICs to justify the adoption of M3D integration in the advanced technology nodes. Then, we present the unique physical design solution named Compact-2D flow, which produces commercial-quality gate-level 3D IC layouts. Finally, we adopt the LSM architecture to build an online machine-learning hardware platform, and study the power-performance-area benefits of 3D ICs to the non-trivial speech recognition application.

1.4 Organization and Contributions

Each research is organized into a self-contained chapter, and the key contributions of this dissertation are as follows:

- In Chapter 2, a new layout method, the stitching scheme, is proposed to improve cell performance and power integrity of T-M3D ICs. Compared to 2D ICs at iso-performance, stitching T-M3D ICs show a maximum of 6% power savings, 44% area savings with only 1% more static IR-drop in the 14nm technology node while existing T-M3D designs undergo serious degradation in static power integrity, causing a reliability issue.
- In Chapter 3, we develop highly-accurate full-chip, GDSII-based wafer and die cost model for 2D and M3D ICs, and inspect the cost overhead issue in G-M3D ICs in detail. To further improve the area savings of G-M3D ICs, a new physical design methodology named Projected-2D flow is developed, and we study how much cost should be further reduced to justify the adoption of M3D technology at the 7nm era.
- In Chapter 4, a physical design solution named Derated-2D flow for G-M3D ICs to tackle inter-tier FEOL/BEOL degradation is proposed. Using a 7nm bulk FinFET from a foundry-grade process design kit (PDK), we model the mobility degradation of the top tier device caused by the low thermal budget process, and quantify the impact of both W BEOL and cost-driven metal layer savings in the bottom tier on M3D design performance. Experiments show that the Derated-2D allows only 3% performance degradation in G-M3D ICs under the worst FEOL/BEOL degradation scenario.
- In Chapter 5, we propose a full-chip RTL-to-GDSII physical design solution to build high-density and commercial-quality two-tier F2F-bonded 3D ICs. The state-of-the-art flow named Shrunk-2D (S2D) [26] requires shrinking of standard cells and interconnects by a factor of 50% to fit into the target 3D footprint of a two-tier design. This, unfortunately, necessitates commercial place/route engines that handle one node smaller geometries, which can be challenging and costly. Our flow named Compact-2D (C2D) does not

require any geometry shrinking. Instead, C2D implements a 2D IC with scaled interconnect RC parasitics, and contracts the layout to the F2F design footprint. In addition, C2D offers post-tier-partitioning optimization that is shown to be effective in fixing timing violations caused by inter-tier 3D routing, which is completely missing in S2D. Lastly, we present a methodology to recycle the routing result of post-tier-partitioning optimization for final GDSII generation. Our experimental results show that at iso-performance, C2D offers up to 26.8% power reduction and 15.6% silicon area savings over commercial 2D ICs without any routing resource overhead.

- In Chapter 6, the design and architectural co-optimization of hierarchical M3D LSM-based neuromorphic processor is studied. By utilizing shared memory and adjusting the synaptic model complexity, as well as maximizing M3D IC benefits, we achieve up to 70.0% reduction in the power-performance-area-accuracy overhead for the non-trivial speech recognition task.
- In Chapter 7, we thoroughly analyze the impact of the reservoir size and the connectivity density on the classification accuracy and their power-area overhead of the 2D LSM processor designs. Then, we investigate the area-power benefits in the two-tier F2F-bonded 3D LSM processor design using Compact-2D flow. We observe that F2F-bonded 3D integration brings us 52% form factor savings, and additional power savings while preserving the 92% classification accuracy.
- In Chapter 8, we summarize each work and present the future directions of this research.

CHAPTER 2

TRANSISTOR-LEVEL MONOLITHIC 3D STANDARD CELL LAYOUT OPTIMIZATION FOR FULL-CHIP STATIC POWER INTEGRITY

2.1 Transistor-level Monolithic 3D Integration

Depending on the granularity of 3D interconnections, M3D designs are categorized into the block-level (B-M3D), the gate-level (G-M3D), and the transistor-level (T-M3D). While B-M3D and G-M3D use MIVs to route the 3D nets of blocks or gates placed on different tiers, T-M3D uses ultra-dense MIVs inside standard cell designs to connect transistors on separate tiers [27, 28]. One of the major challenges in adopting M3D technology is the high manufacturing cost of multiple device layers and metal stacks [29]. To decrease the cost of M3D wafers, industry must reduce the number of metal layers to the maximum allowable extent. While B-M3D and G-M3D need global and intermediate metal resources for both tiers, T-M3D requires only local interconnects on the bottom tier. Therefore, T-M3D is the more favorable solution from an economic perspective.

Another challenge is to maintain the performance of devices and the integrity of interconnects on the bottom tier during the fabrication process. To minimize the impact of post-thermal exposure on the bottom tier, the maximum manufacturing thermal budget for the top tier is constrained under 450 °C [30]. In T-M3D, implementing thermally stable devices on the bottom tier is an additional option to controlling variation since T-M3D allows us to place NMOS and PMOS on separate tiers. According to experimental results in [15, 30], the active sheet resistance of NMOS is more susceptible to post-thermal exposure than that of PMOS because of doping deactivation in high concentrations. As a result, existing studies [27, 28] place the pull-down network (PDN) on top of the pull-up network (PUN) in T-M3D standard cell designs. However, T-M3D requires special effort to create efficient

standard cell layouts, and does not guarantee more than 50% footprint savings at the full-chip level since each T-M3D cells cannot be 50% smaller than 2D cells due to the MIV spacing rules.

In [27], T-M3D standard cell layouts of the 45nm technology have been designed based on the folding scheme, which retains the same routing topology as 2D layouts while simply folding and placing the PDN on top of the PUN, shown in Figure 2.1. Although the folding scheme reduces the effort at creating T-M3D standard cell layouts, it leaves a huge margin for layout optimization in T-M3D standard cells. In addition, the ground rail overlaps the power rail, resulting in restricted power connections to each standard cell. More recent study [28] shows the impact of optimized local interconnections in the folding T-M3D cell layouts using the 14nm technology by eliminating dummy poly regions that were originally required in 2D layouts, but they still do not address the power delivery issue.

In this research, we propose a new T-M3D standard cell layout optimization method, the stitching scheme, targeting towards improvements in static power integrity. While folding T-M3D cells use only one MIV channel reserved for 3D routing at the bottom side of layouts, the stitching scheme allows two MIV channels at the top and bottom sides at the expense of a small extension of the cell height, and exposes the power and ground rails on the top tier over the MIV tracks. Based on extensive analysis with our 14nm T-M3D technology process design kit, we show that stitching T-M3D cells reduce the wirelength and parasitics for local interconnection because of their two MIV channels. Also, we present that the stitching scheme addresses the intrinsic static power integrity issue of the folding scheme at the full-chip level.

2.2 14nm T-M3D Technology Process Design Kit Development

This section presents the development of our 14nm T-M3D technology process design kit. Figure 2.2 shows technology parameters for the device region of the 14nm T-M3D technology used in this work. Due to the availability, we customize parameter values derived from

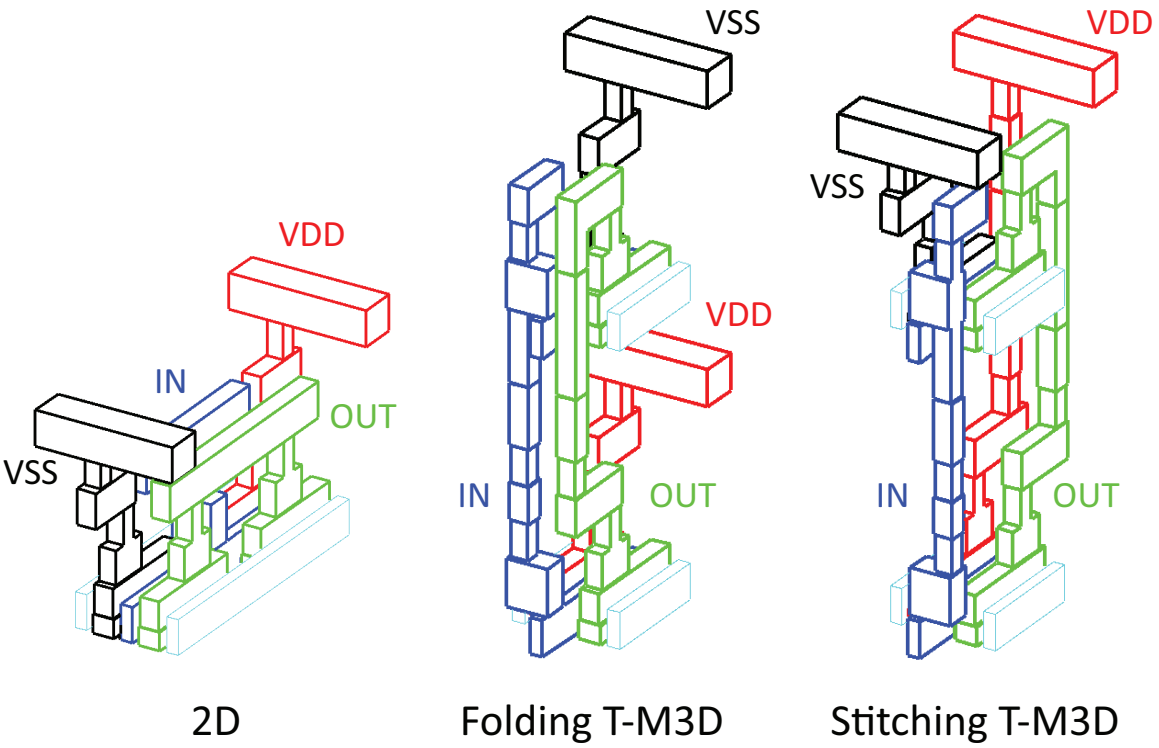


Figure 2.1: 2D and T-M3D INV_X1 cell layouts. To create a T-M3D layout, the folding scheme simply folds the pull-down network and places it on top of the pull-up network while retaining the routing topology of a 2D layout. However, The stitching scheme utilizes two MIV tracks on the top and bottom sides of the layout underneath the power and ground rails. This optimizes internal parasitics and improves static power integrity by exposing both VDD and VSS rails to the back end of line directly.

open-source 14/16nm Predictive Technology Model (ASU-PTM-MG-HP) [31], 2013 International Technology Roadmap for Semiconductors (ITRS) interconnect technology report [32], NCSU FreePDK15 [33], and foundry information available from the public domain [34, 35, 36].

2.2.1 Technology Assumption

We narrow the scope of our research down to the impact of layout optimization in two-tier T-M3D standard cells. Therefore, we assume that device models for the top and bottom tier of T-M3D cells are equivalent to the model of 2D cells. We use 14nm ASU-PTM-MG-HP [31] for the transistor model and employ the middle-of-line (MOL) structure from NCSU

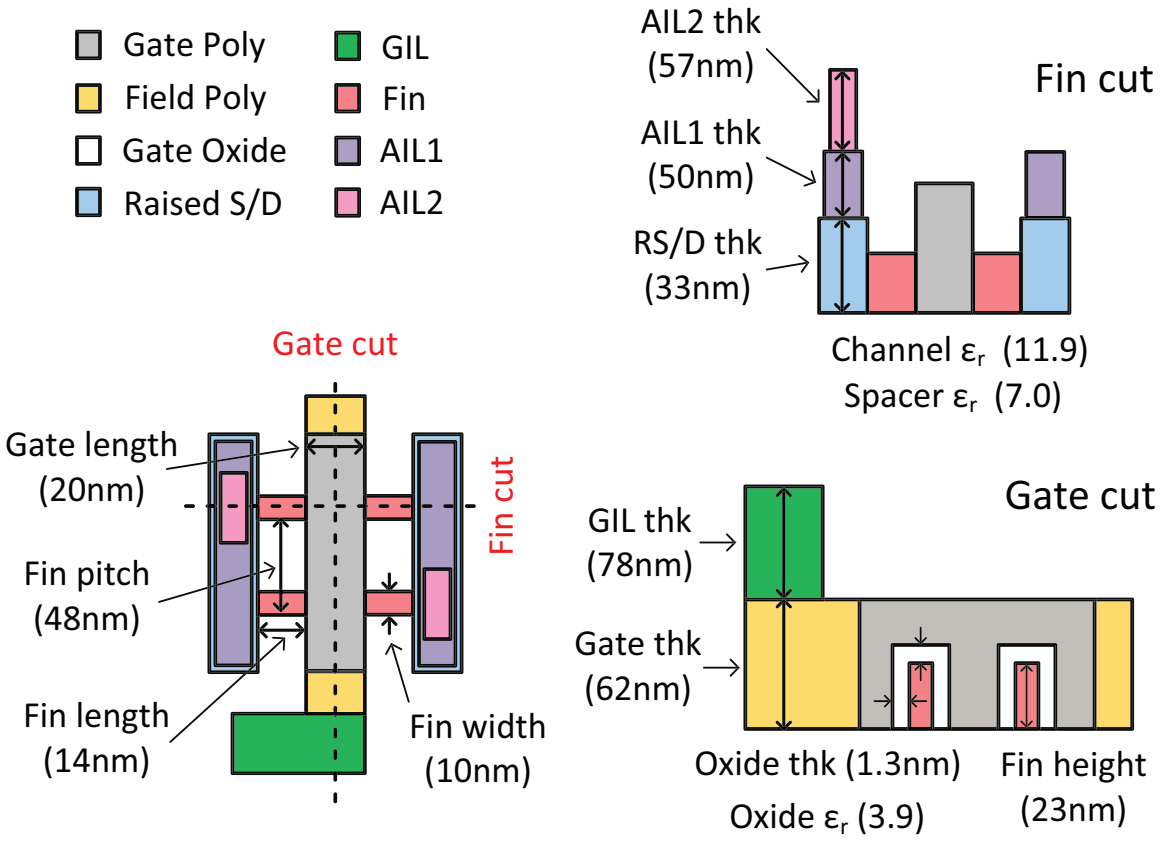


Figure 2.2: Technology parameters customized from [31, 34, 36, 35, 32, 33, 37] for the device region of the 14nm T-M3D technology used in this work. GIL stands for a gate interconnect layer and AIL for a active interconnect layer that correspond to the middle-of-line (MOL) layers from [33]. The contacted poly pitch is 80nm, and the sheet resistances of a gate poly and a raised source/drain are 11.0 Ω/sq and 13.0 Ω/sq, respectively. The resistivity of each MOL layer is 0.07 $\Omega \cdot \mu m$, and the supply voltage is 0.8V. The same parameters are assumed for both top and bottom tiers. Thk in the figure indicates the thickness.

FreePDK15 [33] for both top and bottom tiers. The MOL structure has a gate interconnect layer (GIL) for the gate contact, an active interconnect layer-1 (AIL1) for the connection between individual fins of a device, and an active interconnect layer-2 (AIL2) between a AIL1 and a Metal1 (M1) [33]. Both the GIL and AIL2 layers have a connection to the M1 layer through the Via0 layer (V0). The thickness of each MOL layer is modified from the original structure in FreePDK15 to reflect the device parameters of our 14nm T-M3D technology, but the total height of the MOL structure is retained. We provide T-M3D cells

with two local routing layers on the bottom tier (M1B, M2B) and the top tier (M1T, M2T). The M1 pitch is 80*nm*, which is the same as the contacted poly pitch (CPP), and the M2 pitch is 64*nm*. Both M1 and M2 are 60*nm* thick, and the aspect ratio is 1.875. The low-K dielectric constant (ε_r = 2.55) and the conductor resistivity ($\rho = 0.0451\Omega \cdot \mu m$) are both derived from [32]. Since the inter-layer dielectric (ILD) should electrically separate device regions on the top tier from closely spaced interconnect lines on the bottom tier, we assume that the thickness of the ILD layer is 100*nm* to prevent the threshold voltage of devices on the top tier from changing over 5% [37].

An MIV should make a connection between the top and bottom tier without significant area overhead. Previous studies [27] used the 45nm planar CMOS technology, which allows the insertion of MIVs without increasing the height of standard cells since PMOS is larger than NMOS. However, the 14nm FinFET CMOS technology requires the same number of fins in PMOS and NMOS, and the areas of the devices are also the same. If an MIV makes a connection only between an M2B and an M1T layer, we must extend the height of T-M3D standard cells because of the spacing rule between the MIV and the GIL layer on the top tier (GILT), or between the MIV and MIVs in neighboring cells. An example of the folding T-M3D layout of the INV_X1 cell in Figure 2.3 clearly demonstrates this problem. If we assume that the minimum width and the spacing of an MIV is 32*nm*, the height of T-M3D cells should be six metal tracks (6T) high, resulting in only 33% of area savings in T-M3D cells over 9T 2D layouts. Moreover, the wirelength of local interconnects is lengthened because 3D routing detours the device region on the top tier, degrading the performance and power savings of T-M3D cells. Therefore, we assume that an MIV can be fabricated underneath a GILT layer, in which MIVs directly connect an M2B and the GILT layer while not penetrating the active regions of the top tier. Now, we have two MIVs in our 14nm T-M3D technology. One is an MIVM that connects an M2B and an M1T, and the other is an MIVG that connects an M2B and a GILT. With these two MIVs, folding T-M3D cells achieve 44% of area savings over 9T 2D cells. Taking the ILD thickness (100*nm*) into

account, the aspect ratio of an MIVG is 5, and 7.5 for an MIVM.

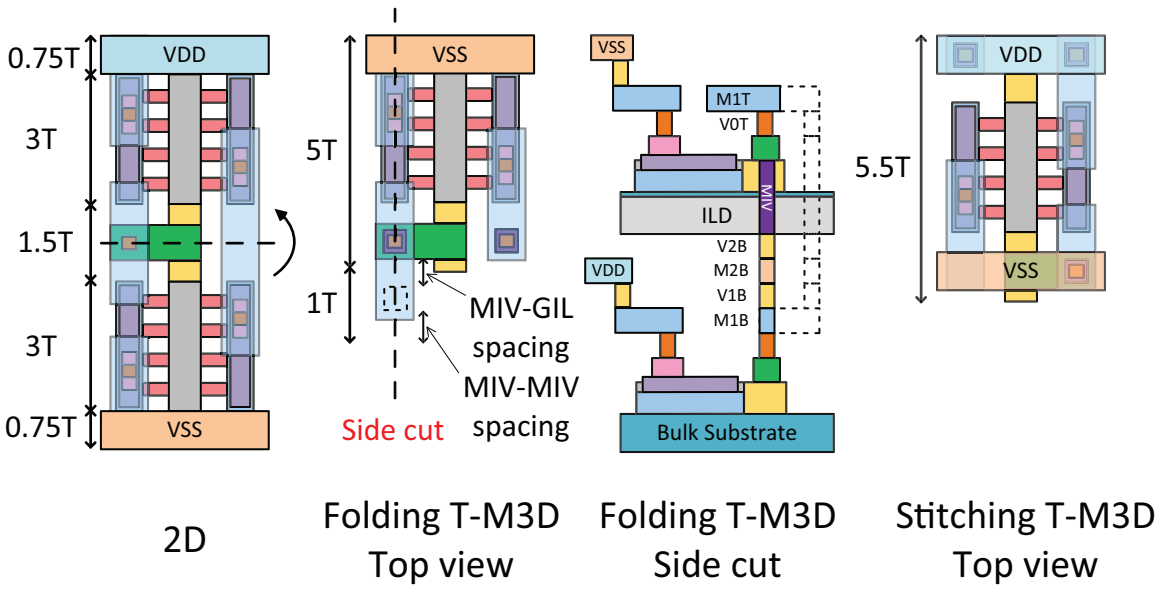


Figure 2.3: An example of the folding and stitching T-M3D INV_X1 cell layout to show the requirement of an additional assumption on the MIV layer in the 14nm T-M3D technology. If an MIV makes a connection only between an M2B and an M1T layer, we must extend the height of T-M3D standard cells, resulting in the degradation of area and performance savings in T-M3D cells. Therefore, in our 14nm T-M3D technology, we assume that an MIV can be fabricated underneatch a GILT layer, in which MIVs directly connect an M2B and the GILT layer while not penetrating the active regions of the top tier.

2.2.2 Technology Characterization

Characterization flow starts from creating a technology file (.tf) that defines every layer in the developed T-M3D technology. Based on the layer definition of technology file, we create T-M3D standard cell layouts in the GDS format, and carry out LVS and DRC check with rule files compatible with our 14nm T-M3D technology. The LVS and DRC rule files are modified from FreePDK15 [33], and DRC check is done on the top and the bottom tier independently. Next, we prepare for interconnect files (.ict, .itf) that describe all technology parameters presented in the previous section, and then transform these files into the parasitic database (.tch, .nxtgrd). Since T-M3D layout architecture presents vertically coupled situations that are unpredicted in the 2D extraction rule, we leverage a field-solver

engine in the commercial tool to accurately calculate the parasitics of T-M3D standard cell layouts. We also utilize FinFET modeling capability of the commercial extraction engine in this step. However, since the tool is designed for 2D ICs, FinFET modeling for the top tier device is not available. Therefore, we manually ignore the double-counted internal parasitics of device layers that the transistor model already contains. The extraction results generate SPICE netlists with all the parasitics (.spf), and the netlist is used for modeling timing/power of T-M3D cells (.lib, .db). Finally, we abstract the cell layout in the layout exchange format (.lef), and complete our 14nm T-M3D technology PDK generation flow. Figure 2.4 summarizes the overall generation flow.

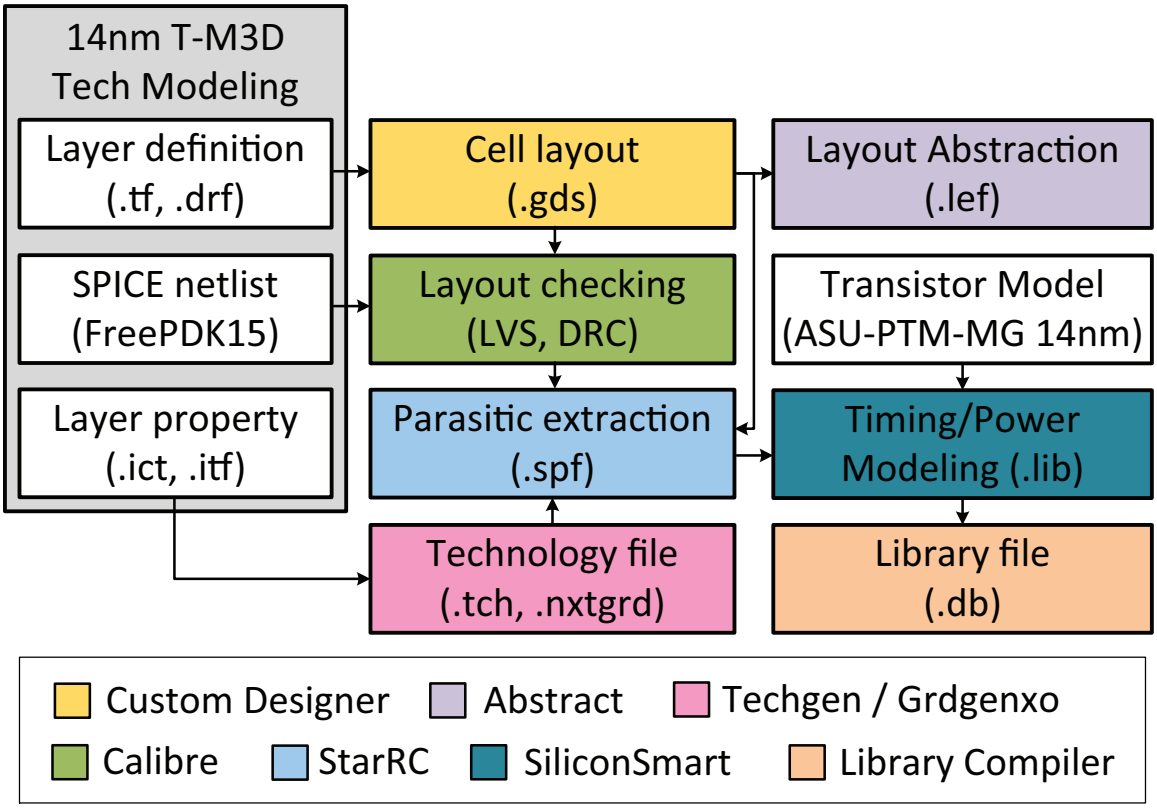


Figure 2.4: PDK generation flow of our 14nm T-M3D technology.

2.2.3 14nm 2D Standard Cell Library

Although 15*nm* 2D Open Cell Library (Nangate 15) [38] is available, cell layouts are twelve metal tracks (12T) high. The number of fins for each transistor is seven, and the CPP is 64*nm*, which is not compatible with the industry standard [28, 35, 39]. We create our own 14*nm* 2D standard cell library targeting 9T cell height with four fins per device. This makes them compatible with the industry-grade 2D layouts which use the CPP as 80*nm* and the metal pitch as 64*nm*. The template-based unidirectional routing scheme [40] is used in the cell layouts. We use 11 routing templates to create our 41 standard cells: (INV, BUF)_X(1, 2, 4, 8), (NAND, NOR, AND, OR) (2, 3, 4)_X(1, 2), (AOI, OAI) (21, 22)_X(1, 2), DFFRNQ_X1. Our 2D layouts are composed of 3T for each device region, 0.75T for the power / ground rails, and 1.5T for the gate contact region.

2.3 Impact of T-M3D Cell Layout Scheme

This section analyzes advantages of the stitching scheme over the folding scheme in the T-M3D standard cell layouts.

2.3.1 Footprint Analysis

While the folding scheme provides T-M3D cells with only one MIV channel at the bottom side of the layouts, the stitching scheme allows two MIV channels at the top and the bottom edge of the layouts. Adding one more MIV channel in the stitching T-M3D cell layouts leads to the extension of cell height by 0.5T compared with folding T-M3D cells. This is under the assumption that we have to honor the minimum width and spacing rule of the MIV layer to prevent overlap with neighboring cells and for reliable MIV fabrication. As a result, folding and stitching T-M3D cells are 5T and 5.5T high, turning into 44.4% and 38.8% of the height of 2D 9T layout, respectively.

Although stitching T-M3D cells are taller than folding T-M3D cells, we should account

for the impact of each scheme on the width of cell layouts to evaluate the final footprint savings. This is because two MIV channels in stitching T-M3D cells simplify the routing topology in the large standard cells that suffer from complex local interconnection. For example, compared with the 2D and the folding T-M3D layouts, the stitching T-M3D D flip-flop (DFF) layout decreases the width of the cell by 320*nm* (12.5%) shown in Figure 2.5. Therefore, the stitching T-M3D DFF achieves 4% lower layout footprint than that of the folding T-M3D DFF, and 52% lower than that of the 2D DFF.

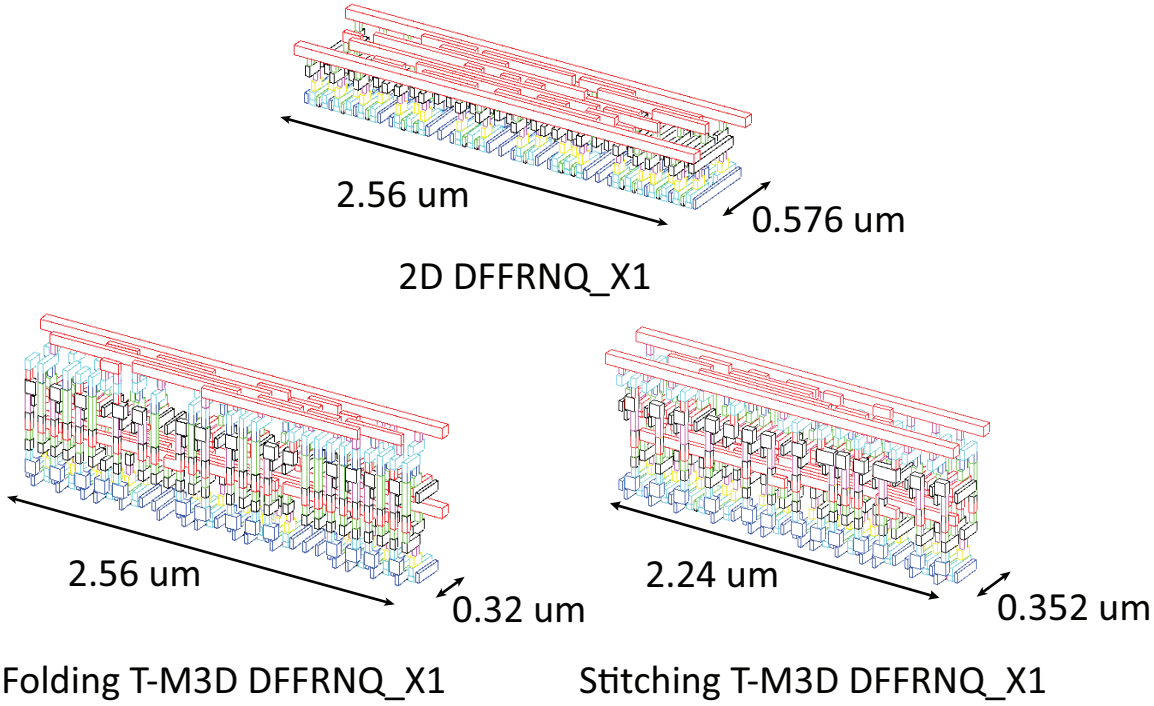


Figure 2.5: An example of a DFF cell with 2D, folding T-M3D, stitching T-M3D layouts. Compared with the 2D and the folding T-M3D counterpart, well-designed stitching T-M3D D flip-flop (DFF) layout reduces the cell width by 320*nm* (12.5%), resulting in 4% of more layout footprint savings over folding T-M3D DFF, and 52% of savings over 2D DFF.

As we observe in the example of a DFF cell, the footprint savings of stitching T-M3D cells depend on the level of routing optimization using two MIV channels. Folding T-M3D cells with simple internal routing has better footprint savings than stitching T-M3D cells, but the large and complex T-M3D cells achieve the footprint savings from the stitching scheme because of two MIV channels since they have much room for the routing optimiza-

tion.

2.3.2 Parasitic Analysis

We analyze the parasitics of T-M3D layouts based on the extraction results of the simplest layout. Table 2.1 shows net coupling capacitances in the INV_X1 layout of the 2D, folding T-M3D, and stitching T-M3D cells. We first address the noticeable changes in the parasitics of the folding T-M3D layout. Compared to the coupling capacitance between the IN and OUT net of the 2D parasitics, that of the folding T-M3D INV_X1 layout increases 45.4% (49aF). After all, like the 2D layout, the folding scheme retains the routing of the IN and OUT nets in parallel, while vertical interconnection with an MIV lengthens the routing interconnects of the IN and OUT nets. Both the IN and OUT nets in the folding T-M3D layout go through every metal stack, including M1B, V1B, M2B, V2B, (MIVG, MIVM), (GILT, M1T), V0T, and M1T. Therefore, the sum of area facing each other between the IN and OUT nets in the folding T-M3D layout is $0.023\mu m^2$ while it is $0.017\mu m^2$ for M1 of the 2D layout. Analysis of layer-by-layer capacitance clearly explains the increase in coupling capacitance between the IN and OUT nets. Coupling between layers in the parallel pillars for the IN and OUT nets contribute to 21aF capacitance. The GILT layer of the IN net and the MIVM layer of the OUT net are also very close ($12nm$) because of the one MIV channel in the folding scheme. Hence, coupling between these two layers contributes another 10aF capacitance. A decrease of 81.3% (26aF) in the VDD and VSS net coupling is noteworthy because the VDD and VSS nets do not face each other in the folding T-M3D layout, and ILD electrically separates the coupling between the top and bottom tiers.

With regard to the parasitics of the stitching T-M3D layout, we observe only an 8% (9aF) increase in coupling capacitance between the IN and OUT nets. The stitching T-M3D layout does not suffer from huge coupling capacitance between the IN and OUT nets. This is because each net uses an independent MIV channel. However, a small increase comes from the short distance between the MIVG layer of the IN net and the diffusion

Table 2.1: Comparison of net coupling capacitance for INV_X1. Compared to the 2D layout, the coupling capacitance between the IN and OUT nets increases 45.4% in the folding T-M3D layout while it is only an 8% increase in the stitching T-M3D layout.

2D (fF)	IN	OUT	VDD	VSS
IN	-	0.108	0.057	0.058
OUT	0.108	-	0.027	0.027
VDD	0.057	0.027	-	0.032
VSS	0.058	0.027	0.032	-
Folding T-M3D (fF)	IN	OUT	VDD	VSS
IN	-	**0.157**	0.055	0.061
OUT	**0.157**	-	0.028	0.022
VDD	0.055	0.028	-	**0.006**
VSS	0.061	0.022	**0.006**	-
Stitching T-M3D (fF)	IN	OUT	VDD	VSS
IN	-	**0.117**	0.056	0.064
OUT	**0.117**	-	**0.058**	0.019
VDD	0.056	**0.058**	-	0.022
VSS	0.064	0.019	0.022	-

layer of the OUT net. Since the VDD and OUT nets share the same MIV channel in the stitching T-M3D layout, coupling between the parallel pillars for the OUT and VDD nets result in a 107.1% (30aF) increase in coupling capacitance between the VDD and OUT nets.

The long wirelength of the vertical interconnections with MIVs also significantly impact the ground capacitance and the resistance of the 3D net. In the 2D layout, the ground capacitance of the IN net is only 4aF, but the ground capacitances of the IN net in the folding and stitching T-M3D are 18aF and 17.2aF, respectively. On an average, the resistance values of the IN and OUT nets of the folding and stitching T-M3D layouts are 13% more than that of the 2D layout. An increase in the resistance of the VDD net in the stitching T-M3D layout is noticeable. While the resistance of the VDD net in the 2D layout is 15.4Ω, it becomes 43.4Ω in the stitching T-M3D layout.

In summary, parasitic analysis shows that T-M3D layouts suffer from additional parasitics. They do not fully turn the huge footprint savings into the power/performance savings. This is because the wirelength of a net that becomes 3D by using vertical in-

terconnection with MIVs is longer than that of the net in the 2D layout. However, T-M3D layouts can improve parasitics when the net in the 2D layout is long enough for routing optimization using multiple MIVs. This implies that the vertical dimension should be furthur reduced by minimizing the metal and ILD thickness to maximize the benefits from T-M3D integration.

2.3.3 Cell Power and Performance Analysis

Based on the results of timing/power model (.lib) of the 41 standard cells of the 2D, folding T-M3D, and stitching T-M3D layouts, Table 2.2 shows the best, worst, and average savings in the timing and power metrics of T-M3D cells normalized to 2D metrics. On an average, folding T-M3D cells show a small degradation in both delay and power consumption compared to 2D cells. On the other hand, The stitching T-M3D cells show benefits in power consumption. In terms of the best timing and power savings compared to the 2D metrics, the stitching T-M3D NOR4_X1 cell reduces power by 6.12% while the folding T-M3D NAND4_X2 cell reduces power by 4.44%. The improvement ratio in the timing metrics of stitching T-M3D layouts is also higher than that of folding T-M3D layouts. Although the stitching scheme has a disadvantage in footprint savings, its two MIV channels reduce coupling between vertical routings. This leads to better timing and power than the folding scheme.

2.4 PDN Design Methodology for Folding T-M3D ICs

In the stitching T-M3D cell layout, VDD and VSS rails are located on the top and bottom edges of the cell boundary in the same way of the traditional 2D layout. Therefore, full-chip T-M3D ICs with the stitching T-M3D standard cells (Stitching T-M3D ICs) are designed with existing 2D CAD engines without any problems. However, full-chip T-M3D ICs with the folding T-M3D standard cells (Folding T-M3D ICs) are in need for a novel power delivery network (PDN) design methodology. The reason is that the ground and

Table 2.2: The best, worst, and average savings in the timing and power metrics of T-M3D cells normalized to the 2D metrics. Δ% indicates the savings compared with 2D. In the best cases, stitching T-M3D cells outperform the folding T-M3D cells.

Folding	Best Cell	Δ%	Worst Cell	Δ%	Ave (Δ%)
Fall Delay	NOR3_X1	0.50	AND2_X2	-5.11	-1.68
Rise Delay	NOR3_X1	0.37	AND2_X2	-5.74	-1.82
Fall Slew	INV_X8	0.78	BUF_X8	-4.62	-1.61
Rise Slew	NOR3_X1	0.14	BUF_X8	-7.17	-2.06
Fall Power	NAND4_X2	4.44	AND2_X2	-10.31	-0.72
Rise Power	NOR3_X1	1.33	OR3_X2	-9.20	-4.03
Stitching	Best Cell	Δ%	Worst Cell	Δ%	Ave (Δ%)
Fall Delay	NOR4_X2	1.97	BUF_X8	-3.84	-0.21
Rise Delay	NOR4_X1	1.40	BUF_X8	-5.39	-1.20
Fall Slew	NOR4_X2	1.09	BUF_X8	-6.42	-0.84
Rise Slew	NOR4_X1	1.09	BUF_X8	-10.60	-2.31
Fall Power	OR4_X1	3.17	AND2_X2	-6.17	-0.84
Rise Power	NOR4_X1	6.12	AND2_X2	-9.31	1.21

power rails overlap in the folding T-M3D cell layout. To evaluate the impact of the T-M3D layout scheme on the full-chip static power integrity, this section presents a PDN design methodology for the folding T-M3D ICs.

When the power is delivered from the VDD ring around the folding T-M3D ICs as proposed in [28], it does not guarantee the tolerable static IR-drop at the center of design. Therefore, we periodically make VDD landing spaces by disconnecting the top VSS rails. After we finish the floorplanning and route bottom VDD stripes only, we create placement and routing blockages underneath the expected VDD landing sites. Placement blockages keep the space empty during subsequent design stages, so that VDD landing sites are not occupied by standard cells. Next, we finish the placement step, and route the VSS stripes. Due to the routing blockages that we have created before, ground rails are disjoint. Then we modify the initial routing blockages to place them only on top of the VSS rails as shown in Figure 2.6(a). This step allows the routing engine to make use of M2 layer effectively during the detail routing step. Now we create the power grid mesh from M3 layer (M3T) in the same way as in 2D ICs. In this stage, VDD is not completely connected but only routing blockages are located at the VDD landing sites. After Clock Tree Synthesis (CTS), routing,

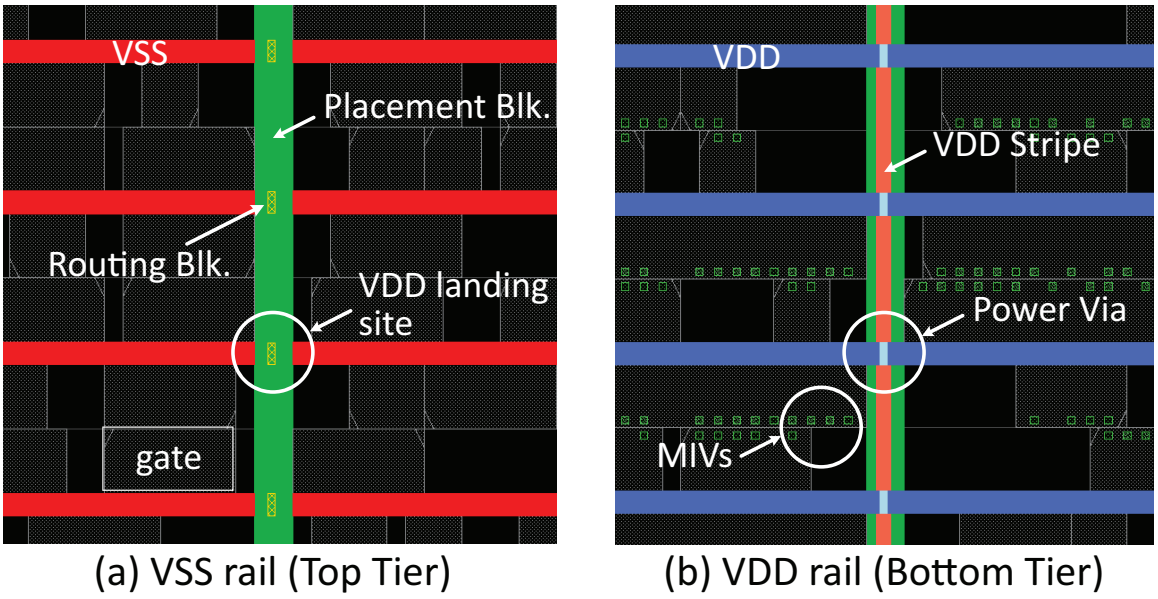


Figure 2.6: Die shots of the PDN in the folding T-M3D IC. (a) Segmented VSS rail routing on the top tier, (b) VDD rail routing on the bottom tier

and post-route optimization are done, we route the bottom VDD rail to connect power grid mesh above the M3T layer while removing the existing routing blockages. Figure 2.6(b) shows the final VDD network on the bottom tier.

2.5 Impact of Full-Chip T-M3D ICs

To evaluate the impact of T-M3D layout optimization on the full-chip design, we choose Triple Data Encryption Standard (DES3), Advanced Encryption Standard (AES), and Low-Density Parity-Check (LDPC) circuit benchmarks from open source hardware benchmark suites [41]. For each benchmark, the core size is set by 70% of the final placement utilization, and the clock period is determined by the worst negative slack which is less than 20ps. We set the clock period of DES3 as 0.4ns, 0.5ns for AES, and 1.2ns for LDPC. PDN metal usage is determined under the condition that the die static IR-drop is less than 2% (16mV) of 0.8V supply voltage for each benchmark. 10% of M5 with 64nm width, 10% of M6 with 128nm width, and 20% of M7, M8 with 384nm width are used for PDN design in

every benchmark. Under the same settings, folding and stitching T-M3D ICs are designed by respective design flows, and voltage sources are placed on every cross-section between M8 and M7 layers. In this work, we do not take the packaging IR-drop into account.

2.5.1 Area and Wirelength Results

Table 2.3 shows the final design result. Area savings in the T-M3D cell layout have a significant impact on the footprint of the full-chip design. It is worth noting that the design footprint savings in the stitching T-M3D ICs vary from 41% to 44%. Since the D flip-flops (DFF) composition in each benchmark varies, the design footprint savings are higher than the cell-level footprint savings (38%). Because the stitching T-M3D DFF cell has simplified routing topology as mentioned in Section 2.3, its area is 4% lower than that of the folding T-M3D DFF cell. Therefore, the more DFF-dominant a circuit is, the better the footprint savings are in the stitching T-M3D designs. DES3 is an example of the DFF-dominant circuit. Out of 57K gates, DFF has 15% of cells as DFF, so stitching T-M3D DES3 design achieves 44% of footprint savings, which is in the comparable level of the footprint savings (45%) from the folding T-M3D ICs.

Wirelength savings follow the huge footprint savings. Folding T-M3D LDPC design achieves 24% lesser wirelength at its best. Since LDPC is a wire-dominant circuit based on the average net length, the impact of the footprint reduction on the wirelength savings is higher than other benchmarks. Stitching T-M3D LDPC has relatively small footprint savings than the folding T-M3D design because only 2% of gates are DFFs. As a result, folding T-M3D LDPC shows more wirelength reduction than the stitching T-M3D design does.

Table 2.3: Full-chip design metric, power, and IR-drop comparison. Each benchmark is timing-closed at the same target for iso-performance comparisons.

Design Style	Footprint (μm^2)	Place. util.(%)	Wirelength (m)	Netlength (μm)	Wire Cap / Pin Cap	Switching Power (mW)	Internal Power (mW)	Total Power (mW)	Static IR-drop (mV)
LDPC circuit, 833MHz									
2D	32585	69.7	1.099	14.07	54:45	68.5	34.9	104.2	11.47
Folding	18020 (-45%)	65.7	0.837 (-24%)	11.28	47:52	62.4 (-9%)	34.0 (-3%)	97.0 (-7%)	62.58 (x5.46)
Stitching	19372 (-41%)	65.01	0.896 (-18%)	12.12	50:49	63.9 (-7%)	33.3 (-5%)	97.9 (-6%)	15.57 (x1.36)
AES circuit, 2.0GHz									
2D	60981	72.4	0.824	6.32	30:70	80.2	93.9	175.5	6.43
Folding	33410 (-45%)	72.2	0.651 (-21%)	5.02	25:75	80.5 (+0%)	95.2 (+1%)	177.1 (+2%)	24.67 (x3.84)
Stitching	35014 (-43%)	72.2	0.671 (-19%)	5.15	26:74	78.7 (-2%)	90.9 (-3%)	170.9 (-4%)	9.43 (x1.47)
DES3 circuit, 2.5GHz									
2D	32596	75.7	0.265	4.63	23:77	61.4	89.0	151.1	8.05
Folding	17789 (-45%)	74.3	0.218 (-18%)	3.91	20:80	64.0 (+4%)	90.5 (+2%)	155.1 (+3%)	44.68 (x5.55)
Stitching	18141 (-44%)	75.2	0.215 (-19%)	3.77	21:79	61.9 (+1%)	87.1 (-2%)	149.7 (-1%)	13.28 (x1.65)

2.5.2 Power and IR-drop Results

Even though the huge wirelength savings result in the wire capacitance savings in T-M3D ICs, the switching power of T-M3D designs shows degradation in AES and DES3. This is because the sum of wire capacitance and gate pin capacitance determines the total switching power of a design. Therefore, the impact of wire capacitance savings on the switching power depends on whether the circuit characteristic is wire-dominant or gate-dominant. Moreover, T-M3D standard cells have larger input and output pin capacitance than 2D on average. For gate-dominant DES3 and AES, wire capacitance reduction is not enough to compensate the pin capacitance increase, resulting in the increase of total capacitance. However, the stitching T-M3D cells reduce the increase of pin capacitance with the help of layout optimization. The switching power derived from pin capacitance is relatively smaller than that of the folding T-M3D designs, and shows more total switching power savings.

Folding T-M3D designs also have more internal power than respective 2D designs due to the degraded cell characteristic. Gate-dominant circuits have more impact from the internal power increase of the folding T-M3D cells. However, stitching T-M3D designs reduce the internal power than 2D in the case of every benchmark. As a result, folding T-M3D designs have 2.5% of total power increase in gate-dominant DES3 and AES, but 7% of savings in wire-dominant LDPC. On the other hand, stitching T-M3D designs always show total power savings. They are 1% of power savings for DES3, 4% for AES, and 6% for LDPC. Folding T-M3D designs have more power savings in the wire-dominant circuit than stitching T-M3D designs do since the 45% of guaranteed footprint reduction leads to more wire capacitance savings.

For the die static IR-drop, folding T-M3D designs do not avoid the huge degradation in the static power integrity because of the limited VDD connections. While folding T-M3D designs show a maximum die IR-drop of 8%, stitching T-M3D designs guarantee less than 2%. Since improving static IR-drop by increasing the PDN metal usage increases

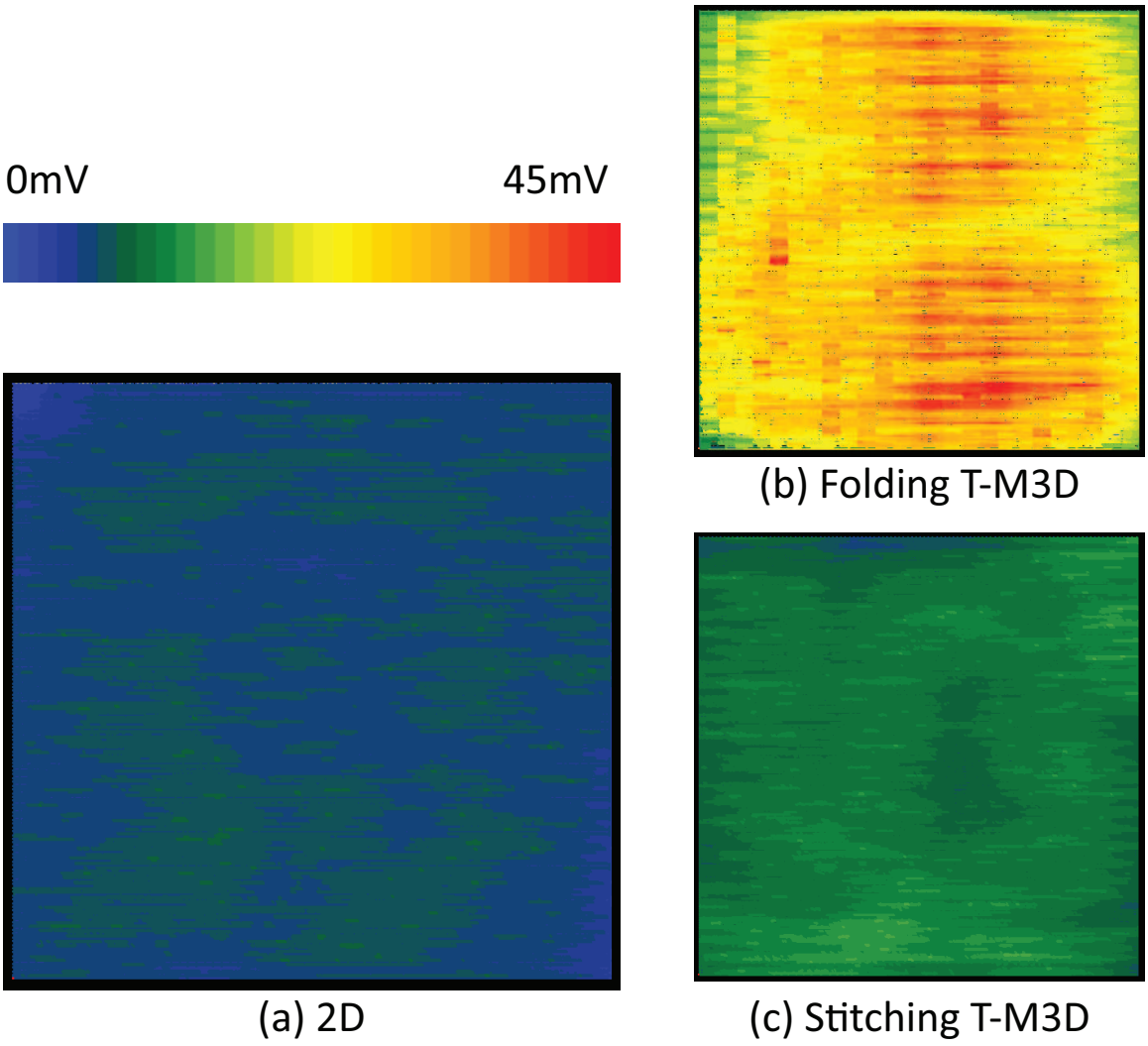


Figure 2.7: Static IR-drop map of DES3. (a) 2D (max = 8.05mV), (b) Folding T-M3D (max = 44.68mV), (c) Stitching T-M3D (max = 13.2mV).

the routing congestion and the number of placement blockages because of the VDD landing sites, the power optimization of the folding T-M3D design is limited.

To summarize, folding T-M3D designs show a maximum footprint savings of 45% and power savings of 7%. However, the severe degradation in static power integrity reduces their reliability. On the other hand, our stitching T-M3D designs guarantee maximum power savings of 6% at static IR-drop less than 2% while achieving design footprint savings greater than 41%.

2.6 Conclusion

In this study, we proposed a new layout optimization method, the stitching scheme, for the transistor-level monolithic 3D (T-M3D) standard cell design. The stitching scheme addresses the static power integrity issue inherent in the folding scheme for T-M3D cell layouts. It also minimizes the timing/power degradation caused by parasitics originating from the unique T-M3D layout architecture. We developed the 14*nm* T-M3D technology process design kit and designed 41 standard cells in the form of 2D, folding T-M3D, and stitching T-M3D layouts. We proved that the stitching scheme outperforms the folding scheme in terms of timing and power metrics at the expense of the increase in the cell height by only 0.5 metal tracks. We also presented a design methodology for a power delivery network in folding T-M3D ICs, and performed sign-off IR-drop analysis in both folding and stitching T-M3D ICs. Lastly, we found that the folding scheme cannot be applied to commercial grade layouts because of its severe IR-drop. However, compared to 2D ICs, the stitching T-M3D ICs experience only 6mV increase in maximum IR-drop while reducing the footprint by up to 44% and power consumption by 6%.

CHAPTER 3

HOW MUCH COST REDUCTION JUSTIFIES THE ADOPTION OF MNOLITHIC 3D IC AT 7NM TECHNOLOGY NODE?

Existing studies on gate-level monolithic 3D (G-M3D) ICs have focused on power, performance, and area improvement in the two-tier design given the same routing resources and silicon area as those of 2D ICs. For example, if a 2D IC has 5 metal layers and $100mm^2$ footprint, then a two-tier M3D IC has 5 metal layers and $50mm^2$ footprint on top and bottom tiers each. Based on those assumptions, [42, 18] shows that G-M3D ICs indeed offer huge iso-performance power savings compared with 2D ICs. Simply and ideally thinking, 50% footprint saving in M3D ICs results in 29.3% wire length reduction ($1/\sqrt{2}$ half perimeter wire length scaling) if the design aspect ratio is assumed to be the same [43]. This wire length savings not only decrease the wire capacitance (switching power savings) but also provides with positive path timing margin to reduce buffer counts (internal power savings). Therefore, if the type of a design is a wire-dominant circuit, power savings in G-M3D ICs are expected to be more.

However, since the footprint of wire-dominant circuits is determined by routability based on the limited routing resources, the design quality of this type of circuit would be easily improved when more routing layers are added. While M3D design needs to have the number of metal layers as few as possible to reduce the fabrication cost, adding more metal layers and optimizing BEOL metal stack in 2D IC can be easily achieved within a reasonable cost overhead [44]. Therefore, it leads to the next questions on how to set the proper 2D reference design for the fair PPC comparison with M3D design, and how much M3D has PPC-competitiveness to make us move toward the M3D era. This chapter addresses above questions.

3.1 Cost Modeling

Previous works [45, 46] on cost modeling for 3D IC are based on estimation of design parameters. Those studies use empirical constant for the area of standard cells, and expected wirelength distribution to predict total die area, and the number of required BEOL layers. In this research, accurate cost models are developed based on the real full-chip GDS design result.

Table 3.1: Nomenclatures for this work.

$C_{W_{FEOL}}$	Manufacturing cost for FEOL		
C_{M_i}	Normalized manufacturing cost for metal layer M_i		
$C_{W_{BEOL,N}}$	Manufacturing cost for N BEOL layers		
$A_{W\|D}$	Wafer | Die area	$Y_{W\|D}$	Wafer | Die yield
D_W	Wafer defect density	DPW	# Dies per wafer
$C_{W\|D_N}$	Wafer | Die cost for 2D IC with N BEOL layers		
$C_{W\|D_{N,M}}$	Wafer | Die cost for M3D IC with N (top) and M (bottom) BEOL layers		
α	Cost variable for M3D top tier manufacturing & bonding		
β	Cost variable for M3D wafer yield degradation		

3.1.1 Wafer Cost Model

Through the cost analysis framework from our industry partner, simple but self-contained wafer cost models are developed for 2D and M3D technology. Considering prescribed sequence of 7nm bulk FinFET process flow, and based on Cost-of-Ownership (CoO) where a database framework considers throughput of fab tools, material, labor, repair, utility and overhead expenses due to the equipment operation [47, 48], the ratio between FEOL and BEOL manufacturing cost is set as 30%:70%. 2D BEOL metal stack configuration used in this research is in accordance with International Technology Roadmap for Semiconductor (ITRS) guidelines for 7nm technology node. Since the foundry-grade 7nm bulk FinFET device technology is assumed to have the middle of line (MOL), MINT layer is included in the metal stack, but it is only used for intra cell routing.

Table 3.2: Assumed patterning option and manufacturing cost per metal layer.

Layer	Patterning	Pitch	Width	Thickness	Normalized Cost (C_{M_i})
MINT (M0)	SAQP	32nm	21nm	24nm	2.8
M1	LELE	42nm	24nm	24nm	1.7
Mx	SAQP	32nm	24nm	24nm	2.8
My	LELE	48nm	24nm	48nm	1.5
Mz	LE	80nm	40nm	80nm	1.0

Table 3.2 shows the assumed patterning option and manufacturing cost per metal layer (C_{M_i}) obtained from our industry partner. Manufacturing costs for MEOL and intermediate interconnect layers are normalized with the cost for global interconnect layer (Mz). With this Table and proposed ratio between FEOL and BEOL manufacturing cost, the reference design is set as 2D IC with 8 BEOL metal layers, and the normalized wafer cost is calculated for another designs with different metal stack as shown below.

$$C_{W_{FEOL}} = 0.3 \times C_{W_8}, C_{W_{BEOL,8}} = 0.7 \times C_{W_8} \tag{3.1}$$

$$C_{W_{BEOL,N}} / C_{w_{BEOL,8}} = \sum_{i=0}^{i=N} C_{M_i} / \sum_{i=0}^{i=8} C_{M_i} \tag{3.2}$$

$$C_{W_N} / C_{W_8} = (C_{W_{FEOL}} + C_{W_{BEOL,N}}) / C_{W_8} \tag{3.3}$$

2D Wafer Cost Model: For N BEOL metal layers,

$$C_{W_N} / C_{W_8} = 0.3 + 0.7 \times \sum_{i=0}^{i=N} C_{M_i} / \sum_{i=0}^{i=8} C_{M_i} \tag{3.4}$$

In literature, no work has previously studied cost estimation for M3D integration. Cost for sequential integration is not fully known yet, and top tier manufacturing should be limited due to the FEOL and BEOL integrity on the bottom tier. Therefore, in this work, it is assumed that the FEOL cost for both tiers are the same as default, and a variable is included to take into account the different device manufacturing cost in each tier and bonding cost (α). M3D BEOL cost is calculated by the sum of BEOL cost for each tier.

M3D Wafer Cost Model: For N (top) and M (bottom) BEOL metal layers,

$$C_{W_{N,M}}/C_{W_8} = 0.6 + \alpha + 0.7 \times (\sum_{i=0}^{i=N} C_{M_i} + \sum_{i=0}^{i=M} C_{M_i}) / \sum_{i=0}^{i=8} C_{M_i} \tag{3.5}$$

3.1.2 Die Cost Model

Considerations for the cost of I/O pins, packaging, testing, and cooling are out of the scope in this work. Assuming that edge clearance and notch height of the wafer are ignorable, the die manufacturing cost takes into account the number of dies per wafer, die yield, and die area. For M3D die yield, sensitivity variable β are multiplied to 2D wafer yield, so that it leads to evaluating how much M3D wafer yield should be improved to guarantee the M3D benefits compared with 2D. Experiments are done with $300mm$ of wafer diameter, and $0.2mm^{-2}$ of D_W, and 0.95 of Y_W. Finally,

2D Die Cost Model: For N BEOL metal layers,

$$DPW_N = A_W/A_{D_N} - \sqrt{2\pi A_W/A_{D_N}} \tag{3.6}$$

$$Y_{D_N} = Y_W \times (1 + A_{D_N} D_W/2)^{-2} \tag{3.7}$$

$$C_{D_N}/C_{D_8} = \frac{C_{W_N}}{C_{W_8}} \times \left(\frac{DPW_8 \times Y_{D_8}}{DPW_N \times Y_{D_N}} \right) \tag{3.8}$$

M3D Die Cost Model: For N (top) and M (bottom) BEOL metal layers,

$$DPW_{N,M} = A_W/A_{D_{N,M}} - \sqrt{2\pi A_W/A_{D_{N,M}}} \tag{3.9}$$

$$Y_{D_{N,M}} = \beta \times Y_W \times (1 + A_{D_{N,M}} D_W/2)^{-2} \tag{3.10}$$

$$C_{D_{N,M}}/C_{D_8} = \frac{C_{W_{N,M}}}{C_{W_8}} \times \left(\frac{DPW_8 \times Y_{D_8}}{DPW_{N,M} \times Y_{D_{N,M}}} \right) \tag{3.11}$$

3.2 Physical Design Solutions

In [26], authors present Shrunk-2D flow to build a full-chip G-M3D IC. The idea of this design flow is to manipulate the powerful optimization capability of the commercial tool built for 2D ICs at pseudo-3D design environment where shrunk layout objects are placed and routed in the floorplan with the same dimension as final M3D footprint. For example, assuming 2-tier, G-M3D design with zero silicon area overhead, the footprint of each tier should become 50% of 2D design footprint. For the Shrunk-2D flow, first the floorplan size is fixed as same as the footprint of final M3D, and shrink the geometric dimension of original 2D layout objects to scale by $\sqrt{2}$. Then the area of standard cells become 50% of original cell area, and also the pitch and width of interconnects become 70.7% of the original. Now, the unit-length RC parasitic is also scaled to let commercial router use the original parasitic of interconnects in optimization stages. This scaling procedure is necessary to remove the overlap between standard cells in the shrunk chip footprint, and to obtain reasonable timing optimization by commercial tool.

However, shrinking layout objects is subject to Design-Rule-Violations (DRV) in the complicated standard cell layouts in the advanced technology nodes. Also, scaling RC parasitics of shrunk interconnects to match the parasitics same as the original either is incorrect due to the exaggerated extrapolation of parasitics with internal algorithm in the commercial tool, or requires many efforts to modify the geometric and electrical characteristics in the interconnect files. Furthermore, those layout objects are not reusable in the design with more than 2-tiers. Lastly, Shrunk-2D Flow possibly maximizes the placement utilization of each tier in M3D design, but it does not fully optimize the design in terms of routing utilization since reduced footprint and effectively routed nets in M3D decreases total wirelength. Therefore, a new physical design solution named Projected-2D is proposed for two-tier G-M3D designs. The main idea of this flow is to reuse 2D design itself as a starting point for implementation of M3D design. The overall design steps are shown below.

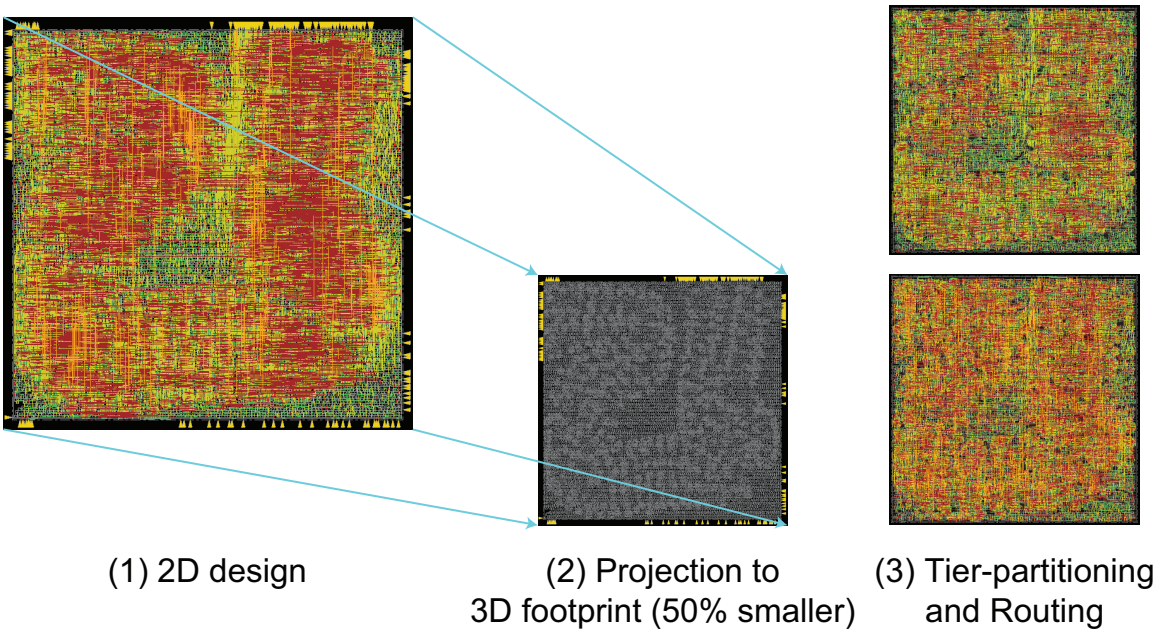


Figure 3.1: Major steps of our Projected-2D flow. (a) 2D IC design, (b) Placement projection, (c) Tier partitioning and tier-by-tier routing after MIVplanning.

3.2.1 Projected-2D Flow

Projected-2D does not require shrinking of layout objects, and scaling RC parasitics unlike Shrunk-2D flow. The beauty of Projected-2D flow is as follows: (1) After 2D design is implemented, which already closes design specification with normal 2D process-design-kit (PDK), Projected-2D reuses final netlist and placement result of the 2D design to implement M3D design. Since there is no difference between the netlist of 2D and that of M3D, it allows to directly compare the routing result of equivalent nets in 2D and M3D designs. Analyzing RC parasitics of those nets allows us to examine the actual wirelength savings in M3D, or to improve tier partitioning result for the better M3D design quality. (2) Projected-2D easily maximizes either placement or routing utilization by projection of 2D placement result. Modulating the projection factor, the final M3D design footprint can be reduced by more than 50% if there is enough routing usage savings. (3) Projected-2D enables multi-tier, gate-level M3D design without any efforts at modifying geometric information in design input files.

However, note that Projected-2D overestimates wire loads, and retains redundant buffers in the 2D design. Table 3.3 shows qualitative, and quantitative comparison between Projected-2D and Shrunk-2D. Assuming 2-tier LDPC M3D design with a foundry-grade 7nm bulk FinFET PDK and 5 metal layers in both tiers, design result of Projected-2D flow has more buffers, resulting in larger positive slack than that of Shrunk-2D. On the other hand, due to the reduced footprint of Projected-2D design, it has more wirelength saving and switching power savings to compromise increase in the internal power caused by redundant buffers.

Table 3.3: Comparison between Projected-2D and Shrunk-2D flow.

	Projected-2D	Shrunk-2D
Shrink macro layout?	No	Yes
Shrink interconnect dimension?	No	Yes
Scale unit-length RC parasitics?	No	Yes
Consider buffer saving in M3D?	No	Yes
Have same netlist as 2D?	Yes	No
Maximize routing utilization?	Yes	No
LDPC M3D result, 7nm bulk FinFET, M5 (top) / M5 (bottom)		
Chip Area (μm^2)	4499	5408
Maximum routing utilization	0.762	0.666
Total buffer count	16163	15980
Total power (mW)	32.76	32.41
WNS (ns)	0.057	-0.015

3.2.2 Tier Partitioning and MIV planning

Based on projected placement location of macros and netlist, placement-driven min-cut partitioning is used for the tier partitioning [43]. This partitioning scheme divides the whole design in regular fashion for the balanced local area skew, so-called partitioning bin, and do Fiduccia Mattheyses (FM) min-cut partitioning inside each of partitioning bins. Therefore, the number of inter-tier connections depends on the size of partitioning bins. In [18], it is shown that there is an optimization point for the minimum power consumption along with the inter-tier connections. This is because too many 3D connections cause routing congestion and redundant snaking between each tiers, while few 3D connections leads to small wirelength savings. Therefore, the best partitioning bin size per benchmark is found

by sweeping the bin size for the maximum power savings.

After tier partitioning, the proper MIV location is decided by using commercial tool built for 2D ICs as proposed in [43]. The main idea of this methodology is to let commercial router treat MIVs as normal vias while there are routing blockages on the area of macros on the top tier to avoid overlaps between MIVs and top macros during routing stage. The limitation of this flow is that the direction of metal layers should not be the same between adjacent tiers, and that the number of interconnect layers on the bottom tier should be an even number. Since our foundry-grade 7nm bulk FinFET standard cell layout contains MINT layer for internal routing, an odd number of interconnect layers on the bottom tier is assumed. Once the MIV locations are determined by MIV planning, a DEF file for each tier is created containing the location of MIVs as primary I/O. Then the timing context of each tier is created to optimize the routing quality. After routing under the timing context, RC parasitics are extracted, and 3D timing and power analysis is proceeded.

3.2.3 Footprint Resizing

Once initial M3D design is done, the maximum placement or routing utilization is checked on each tier if it is over 70%. Since M3D placement utilization on each tier is same as 2D placement utilization considering balanced area skew from placement-driven min-cut partitioning, meeting the sufficient placement utilization is guaranteed from 2D design result. However, if a circuit is BEOL-dominant type, then 2D placement utilization is possible to be lower than 70% because insufficient routing resources requires large die area. In that case, even though 2D routing utilization is over 70% in certain metal layers, routing utilization in M3D could be lower than 70% due to the wirelength reduction. To maximize the utilization of die area, the proper footprint is estimated as $A'_D = A_D \times U_r/0.7$, where U_r is the maximum routing utilization out of all metal layers, A_D is the current footprint area, and A'_D is the updated footprint area. We project the 2D placement into the updated footprint, and iterate the design flow shown in Figure 3.2 until the U_r is over 70%.

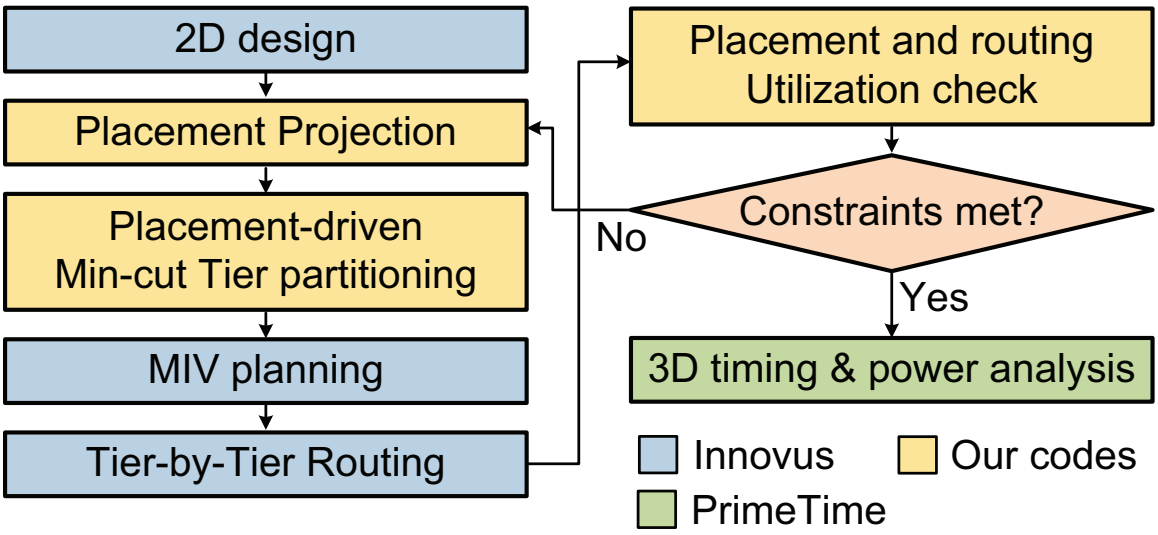


Figure 3.2: Projected-2D design flow.

3.3 Experimental Results

To cover two widely different circuit types, we choose Triple-Data-Encryption-Standard cipher (DES3) and Low-Density Parity-Check decoder (LDPC) circuit benchmarks from open source hardware benchmark suites [41]. 2D Design of these two benchmarks built using a foundry-grade 7nm bulk FinFET PDK shows that 72% of total capacitance in DES3 is pin capacitance while 64% of total capacitance in LDPC is wire capacitance. Also, average net length of LDPC is 3.94 times longer than that of DES3. Therefore, LDPC is defined as a BEOL-dominant circuit, and DES3 as a FEOL-dominant circuit. The diameter and pitch of an MIV in the experiments is assumed to be 24nm and 48nm with resistance of 16Ω and capacitance of $0.01fF$.

Table 3.4: 2D IC PPC analysis and comparisons. Our PPC is defined in Equation 3.12.

Circuit Type	Metal Stack	**Tot. Power (mW)**	Max. Perf (GHz)	Placement Utilization	Max. Routing Utilization	Wafer Cost	Area (μm^2)	DPW (1e+6)	Die Yield	**Die Cost**	**PPC**
FEOL dominant DES3	M3	**37.70**	2.00	0.719	M2, 0.287	0.739	6048	11.679	0.949	**0.739**	**1.306 (best)**
	M4	36.96	2.00	0.718	M3, 0.242	0.804				0.804	1.224
	M5	36.52	1.99	0.716	M3, 0.215	0.870				0.870	1.140
	M6	36.69	2.00	0.716	M3, 0.213	0.913				0.913	1.086
	M7	36.39	1.99	0.716	M3, 0.214	0.957				0.957	1.040
	M8	36.21	1.99	0.715	M3, 0.207	1.000				1.000	1.000
BEOL dominant LDPC	M5	39.28	0.99	0.359	M4, 0.824	0.870	10816	6.529	0.948	1.720	0.433
	M6	33.45	0.99	0.581	M6, 0.807	0.913	6561	10.765	0.949	1.094	0.799
	M7	31.49	0.99	0.686	M6, 0.790	0.957	5476	12.899	0.949	0.957	0.972
	M8	29.28	0.99	0.794	M8, 0.613	1.000	5476	12.899	0.949	1.000	1.000
	M9	28.39	0.99	0.787	M8, 0.678	1.043	4692	15.055	0.949	0.894	1.154
	M10	**27.48**	1.00	0.789	M4, 0.535	1.087	4692	15.055	0.949	**0.931**	**1.156 (best)**

3.3.1 2D Design Results

Table 3.4 shows the impact of changing metal stack configuration on the design result of FEOL-dominant circuit DES3, and BEOL-dominant circuit LDPC. Designs for each benchmark are constrained with the same clock period, (0.5ns for DES3, 1.0ns for LDPC). Total power in the Table 3.4 is iso-performance power number, and the maximum performance is calculated by reversing the sum of clock period and the worst timing slack. PPC is calculated as follows:

$$PPC = \frac{Max\_Performance}{Total\_Power \times Die\_Cost} \tag{3.12}$$

Since wafer and die cost is normalized with that of 8 BEOL metal stack (M8 in Table 3.4) design, PPC is also normalized with the PPC value of M8 design.

FEOL-Dominant Circuit Type

Starting from M8 design, reducing metal layers in FEOL-dominant circuit has little impact on routing utilization overhead. Since most of nets in DES3 is locally routed, maximum routing utilization is only 20.7% in M3 layer even though there are 8 BEOL metal layers for routing. The placement utilization and die area are also unchanged along with metal stack reduction since M3 design already has sufficient routing resources. All designs close the timing, and small change in iso-performance power along with metal stack reduction is caused by slightly increased routing congestion. Even though the total power in M3 design is increased by 4% compared to M8 design, wafer and die costs are reduced by 26%. Therefore, overall PPC saving of M3 design is 31% more than the saving of M8 design, and M3 design is defined as the most optimized design for DES3 in terms of PPC.

BEOL-Dominant Circuit Type

BEOL-dominant circuit LDPC shows interesting results in Table 3.4. In M5 design, the die area is determined by the maximum routing utilization in M4 layer. The lack of routing resources increase chip size even though placement utilization is only 35.9%. The large footprint not only increases die cost, but also makes overall wirelength longer and leads to higher wire capacitance. Therefore, adding only one more metal layer significantly improves the design quality of BEOL-dominant circuit. Compared to M5 design result, M6 design has total power saving by 15%, area reduction by 39%, lower die cost by 36%, and PPC improvement by 85%.

Once there are enough metal layers for the routing in LDPC, the die area needs to be determined by both placement and routing utilization. Therefore, area saving and the impact of adding more interconnect layers become saturated as shown in M8 design. As a result, reduced power saving and additional cost for more metal layer have a tradeoff relationship.

3.3.2 Impact of Metal Stack Optimization

Optimizing dielectric constant, and conductivity in the metal stack by changing material composition is one of the cheapest solutions to improve design quality. We assume that the dielectric constant of global interconnect layers (from M6 to M10) has been reduced by 14%, and generate new technology file (TCH) using Cadence Techgen. Scaling dielectric constant reduces 12% of total capacitance per unit length for the global interconnect metal layers, and this metal stack configuration is defined as Low-K metal stack. We also consider the wafer cost change for the Low-K metal stack. Based on the wafer cost model in Section 3.1, the BEOL cost is increased from 0.70 to 0.71 and takes it into account for the PPC calculation.

Table 3.5 shows the impact of Low-K metal stack on the BEOL-dominant LDPC 2D designs. By comparing M5 design with M5 + Low-K design, reduced wire capacitance by

using Low-K metal stack further improves total power due to the switching power saving. Also, decreased routing congestion from the reduced number and drive strength of buffers make room for die area saving. Since it is assumed that BEOL cost for Low-K metal stack is different from the normal metal stack, it shows different tradeoff between power saving and wafer cost increase. Even though M9 + Low-K design has more power saving than M8 + Low-K design, the PPC value of M8 + Low-K is higher than M9 + Low-K due to the BEOL cost. Overall, M8 + Low-K design is defined as the most optimized design for LDPC with regard to PPC. For the FEOL-dominant DES3 design, the impact of reducing wire capacitance on PPC by using Low-K metal stack is negative since it has little power saving with increased die cost.

Table 3.5: Impact of Low-K metal stack on BEOL-dominant LDPC 2D designs.

Metal Stack	Tot. Power (mW)	Max. Perf (GHz)	Wafer Cost	Area (μm^2)	Die Cost	PPC
M5	39.28	0.99	0.870	10816	1.720	0.433
M5 + Low-K	37.27	0.99	0.878	8190	1.314	0.598
M6	33.45	0.99	0.913	6561	1.094	0.799
M6 + Low-K	32.4	0.99	0.922	6561	1.105	0.818
M7	31.49	0.99	0.957	5476	0.957	0.972
M7 + Low-K	30.72	0.99	0.966	5476	0.966	0.987
M8	29.28	0.99	1.000	5476	1.000	1.000
M8 + Low-K	28.35	0.99	1.010	4692	0.865	**1.194**
M9	28.39	0.99	1.043	4692	0.894	1.154
M9 + Low-K	27.56	1.00	1.054	4692	0.903	1.188
M10	27.48	1.00	1.087	4692	0.931	1.156

3.3.3 M3D Design Results

Table 3.6 shows the M3D design results using normal metal stack of various combinations. 2D design in Table 3.6 is the best design with regard to PPC, defined as the reference for the comparison with M3D design. In this section, the variable for the sequential integration and bonding cost for the top tier (α) is assumed as 0.1, and M3D wafer yield (β) as 90% of 2D wafer yield.

Table 3.6: M3D PPC analysis and comparison. Our PPC is defined in Equation 3.12. Power is total power consumption, and Perf is the maximum performance.

Circuit Type	Design Flavor	Metal Stack (top / bottom)	Power (mW)	Perf (GHz)	Placement Utilization (top / bottom)	Max. Routing Utilization (top / bottom)	Wafer Cost	Area (μm^2)	DPW (1e+6)	Die Yield	Die Cost	PPC
FEOL dominant DES3	2D	M3	37.7	2	0.719	M2, 0.287	0.739	6048	11.679	0.949	0.739	1.306
	M3D	M3 / M5	37.38	1.776	0.744 / 0.718	M2, 0.278 / M4, 0.215	1.826	3041	23.232	0.854	1.019	0.848
		M4 / M5	36.83	1.901	0.745 / 0.718	M3, 0.203 / M4, 0.215	1.872			0.854	1.045	**0.899**
		M5 / M5	36.74	1.901	0.745 / 0.718	M3, 0.187 / M4, 0.215	1.917			0.854	1.070	0.880
		M6 / M5	36.74	1.898	0.745 / 0.718	M3, 0.187 / M4, 0.215	1.948			0.854	1.087	0.864
BEOL dominant LDPC	2D	M8 + Low-K	28.35	0.99	0.794	M8 0.713	1.010	4692	15.055	0.949	0.865	1.194
	M3D	M5 / M5	32.76	1.060	0.481 / 0.425	M4, 0.762 / M4, 0.639	1.917	4499	15.702	0.854	1.750	0.547
		M6 / M5	32.55	1.050	0.563 / 0.491	M6, 0.666 / M4, 0.679	1.948	3894	18.142	0.854	1.538	0.620
		M7 / M5	32.37	1.018	0.563 / 0.491	M6, 0.631 / M4, 0.694	1.978	3894	18.142	0.854	1.562	0.596
		M5 / M7	28.5	1.035	0.606 / 0.528	M4, 0.756 / M4, 0.545	1.978	3504	20.162	0.854	1.406	**0.764**

FEOL-Dominant Circuit Type

While 2D DES3 design with only M3 metal stack already has enough resources to finish the routing, M3D DES3 design should have M5 metal stack in the bottom tier. This is because if M3 metal stack is used in the bottom tier, part of routing resource in M3 layer will be dedicated to inter-tier connection (MIV planning) compromising routability, and leading to many DRVs. Also, due to the limitation of MIV planning scheme using commercial 2D router, the odd number of BEOL metal layers is allowed on the bottom tier so that top metal layer of the bottom tier and MINT layer of the top tier has routing direction orthogonal to each other. Therefore, 5 metal layers are set as the minimum metal stack on the bottom tier, and evaluate the PPC benefit of M3D design.

Since the die area of the FEOL-dominant circuit is determined by placement utilization, 2-tier M3D DES3 design indeed has 50% of footprint savings compared to the 2D design. However, the high wafer cost of M3D integration, and the assumptions on reduced M3D wafer yield increase the die cost for M3D. In addition, total power saving in M3D is not significantly large, since DES3 is FEOL-dominant and most of routing in DES3 are done locally. Performance loss in DES3 M3D design is worth to notice. Because the M3D design keeps the same nets as the 2D design through Projected-2D flow, the worst resistance net in M3D design is compared with the equivalent nets in the 2D design as shown in Table 3.7.

Table 3.7: Equivalent net comparison between M3D and 2D design. The worst resistance net in DES3 M3D design is analyzed.

Wirelength distribution (um)	2D	M3D (top/bottom)
M5	122.35	0.00 / 74.90
M4	67.09	7.42 / 51.74
M3	0.32	5.87 / 3.78
M2	0.27	2.46 / 0.19
M1	0.46	1.50 / 0.46
Net Total Wirelength (μm)	190.50	148.05
Net Total Resistance (Ω)	11187	10206
Unit-length Resistance ($\Omega/\mu m$)	58.72	68.94

It shows detailed wirelength distribution and net resistance of the equivalent net in

the 2D M5 design and the M3D M5 / M5 design. The 2D net has long wirelength, but most of routing are done in M5 layer. However, although the M3D net has 22% total wirelength saving, total net resistance is reduced by 9% only. Unit-length resistance of the M3D net is 17% higher than that of the 2D net. Based on the net comparison, it is observed that when locally placed and routed cells in 2D design are split into different dies through tier partitioning, routing utilizations for intermediate interconnect layers are increased. Since part of the top metal layer in the bottom tier should be dedicated to MIV planning, commercial router is not able to fully use the top metal routing resource in the bottom tier. Instead, it uses more intermediate interconnect layers. Besides, wires should go through the whole metal stack in the bottom tier to route top tier cells. Therefore, it is likely to increase the routing congestion, and redundant wire capacitance.

Furthermore, top tier routing also uses intermediate interconnect layers since only local routing remains. The resistance of M2, M3 layer is 2.46 times higher that of M4, M5 layer. Therefore, locally routed FEOL-dominant circuit requires more effective tier partitioning, otherwise the timing of the critical path worsens. With regard to PPC value, M3D design with M4 / M5 metal stack is defined as the most optimized M3D design for FEOL-dominant DES3.

BEOL-Dominant Circuit Type

When comparing M3D M5 / M5 design with 2D M5 design, BEOL-dominant LDPC M3D design indeed shows an increase of power savings by 17% and die area savings by 58%. However, in Table 3.5, 2D M8 + Low-K design is defined as the reference design for the fair comparison with M3D designs. Since placement and routing utilization of our 2D reference design is highly optimized, die area of 2D design is small enough to offer cheap die cost. Due to the die area as small as 57% of that of the 2D M5 design, huge wirelength and buffer savings result in M3D-compatible power consumption.

Therefore, unlike FEOL-dominant DES3 M3D design, LDPC M3D M5 / M7 design

is found the best M3D design out of given metal stack combinations with regard to PPC value though it only has 25% area saving compared with 2D reference. Table 3.8 shows the impact of Low-K metal stack on LDPC M3D design. By using Low-K metal stack, M5 / M7 + Low-K design finally beats 2D reference in terms of both total power and maximum performance. Even though it is clear that using Low-K metal stack and adding routing resources are very effective solutions to improve M3D design quality, too expensive metal stack for BEOL-dominant circuit increases the wafer cost almost 2 times higher than 2D reference, resulting in lower PPC of M3D than that of 2D.

Table 3.8: Impact of Low-K metal stack on BEOL-dominant LDPC M3D designs.

Metal Stack (top / bottom)	Tot. Power (mW)	Max. Perf (GHz)	Wafer Cost	Area (um)	Die Cost	PPC
2D						
M8 + Low-K	28.35	0.99	1.010	4692	0.865	**1.194**
M3D						
M5 / M5	32.76	1.060	1.917	4499	1.750	0.547
M5 / M5 + Low-K	32.12	1.074	1.929	4499	1.760	0.562
M6 / M5	32.55	1.050	1.948	3894	1.538	0.620
M6 / M5 + Low-K	31.9	1.071	1.960	3894	1.548	0.641
M7 / M5	32.37	1.018	1.978	3894	1.562	0.596
M7 / M5 + Low-K	31.72	1.031	1.991	3894	1.572	0.611
M5 / M7	28.5	1.035	1.978	3504	1.406	0.764
M5 / M7 + Low-K	27.91	1.050	1.991	3504	1.414	**0.787**

3.4 7nm M3D Cost and Yield Study

In Section 3.3.3 and 3.3.3, assuming M3D wafer yield (β) as 90% of 2D wafer yield, and additional cost for top tier device implementation (α) is 10% of wafer cost for the 2D M8 design, it is observed that the PPC of FEOL-dominant DES3 M3D design is worse by 31% and BEOL-dominant LDPC M3D design by 34% compared to the 2D reference. Then the next question is how much M3D wafer yield and additional cost for M3D integration should be further reduced for the cheap M3D die cost to justify the adoption of M3D technology. In Figure 3.3, red surface of each plot shows the valid region along with α, and β where

the best M3D design defined in the previous Sections beats PPC of the 2D reference. Z-axis of these plots is calculated by the ratio of PPC value between M3D and 2D design. We observe that for the adoption of gate-level M3D integration, M3D wafer yield needs to be higher than 90% of 2D wafer yield, and the device manufacturing cost of M3D design should be limited by less than 33% of 2D device manufacturing cost.

Moreover, the experiment result show that FEOL-dominant circuit type has more room for the adoption of M3D, and benefits more from M3D integration than BEOL-dominant circuit type in terms of PPC. This is because the impact of metal stack optimization and giving more routing resources to BEOL-dominant type circuit drastically reduce both power and die area of 2D design compatible to M3D counterpart. The differences in total power and die area between LDPC 2D reference (M8 + Low-K design) and M3D design with the best PPC (M5 / M7 + Low-K design) are only 2% and 25%. However, since the die area of FEOL-dominant circuit type is determined by placement utilization, 50% of footprint saving from M3D technology is guaranteed, resulting in more spaces in terms of die cost for adoption of M3D technology.

Two benchmarks for the previous experiments, DES3 and LDPC, are logic circuits where the number of standard cells in the full-chip 2D design is less than $60k$ based on foundry-grade 7nm bulk FinFET. The chip area of these two small circuits is less than $0.01mm^2$. Since the 2D die yield of those extremely small benchmarks is already sufficient, it explains why the huge footprint saving and die cost benefit from M3D technology does not show up. Therefore, the impact of die area of logic-only design on the die cost of M3D and 2D design is evaluated based on the cost models proposed in Section 3.1. Since the 2D die area of BEOL-dominant circuit is effectively reduced when more routing resources are used, footprint saving of gate-level 2-tier M3D design is only 25% as shown in Section 3.3.3. When the die area is determined by placement utilization like FEOL-dominant circuit, 50% of M3D area saving is guaranteed as analyzed in Section 3.3.3.

We assume that the ratio of the die area of 2D design and that of M3D design is fixed

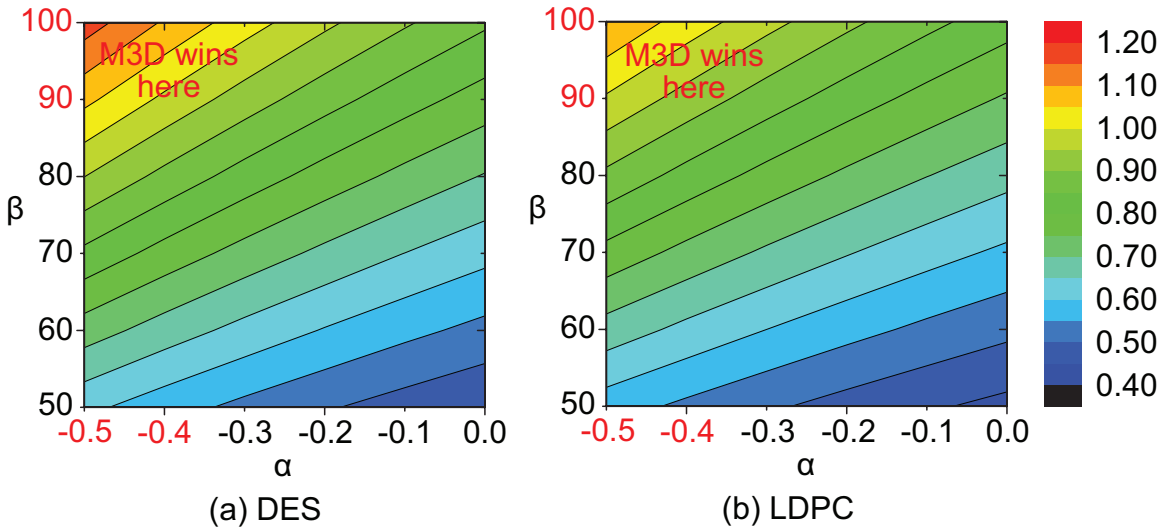


Figure 3.3: M3D cost vs. yield vs. PPC sensitivity analysis. α denotes cost variable for top-tier devices fabriacation and bonding in M3D, e.g., α = -0.4 means that FEOL manufacturing cost for M3D (0.6) should be 67% lower (0.6 + α = 0.2). β denotes M3D wafer yield (percentage w.r.t. 2D wafer yield). Z-axis denotes PPC ratio of M3D over 2D, e.g., 1.2 means M3D PPC is 20% better.

in each circuit type, and calculate die cost for each design scheme considering die yield. Figure 3.4 shows that M3D die cost becomes cheaper than 2D die cost along with the increase in die size. M3D design of FEOL-dominant circuit has significant die cost saving compared to 2D design starting from $2mm^2$ while M3D design of BEOL-dominant circuit becomes cheaper from $70mm^2$ as well. In addition, with the same die size of design for two circuit types, the gap for the ratio between 2D and M3D die cost of FEOL-dominant and BEOL-dominant circuit becomes wider along with die size increase. Assuming $100mm^2$ of 2D die size, FEOL-dominant circuit has 2.5 times more cost competitiveness from M3D technology than BEOL-dominant circuit. The result indicates FEOL-dominant circuit benefits sooner and more from M3D technology in terms of cost than BEOL-dominant circuit.

3.5 Conclusion

This study shows power, performance, and cost (PPC) tradeoffs with full-chip GDS based cost modeling for 2-tier, gate-level, full-chip GDS M3D ICs built using a foundry-grade

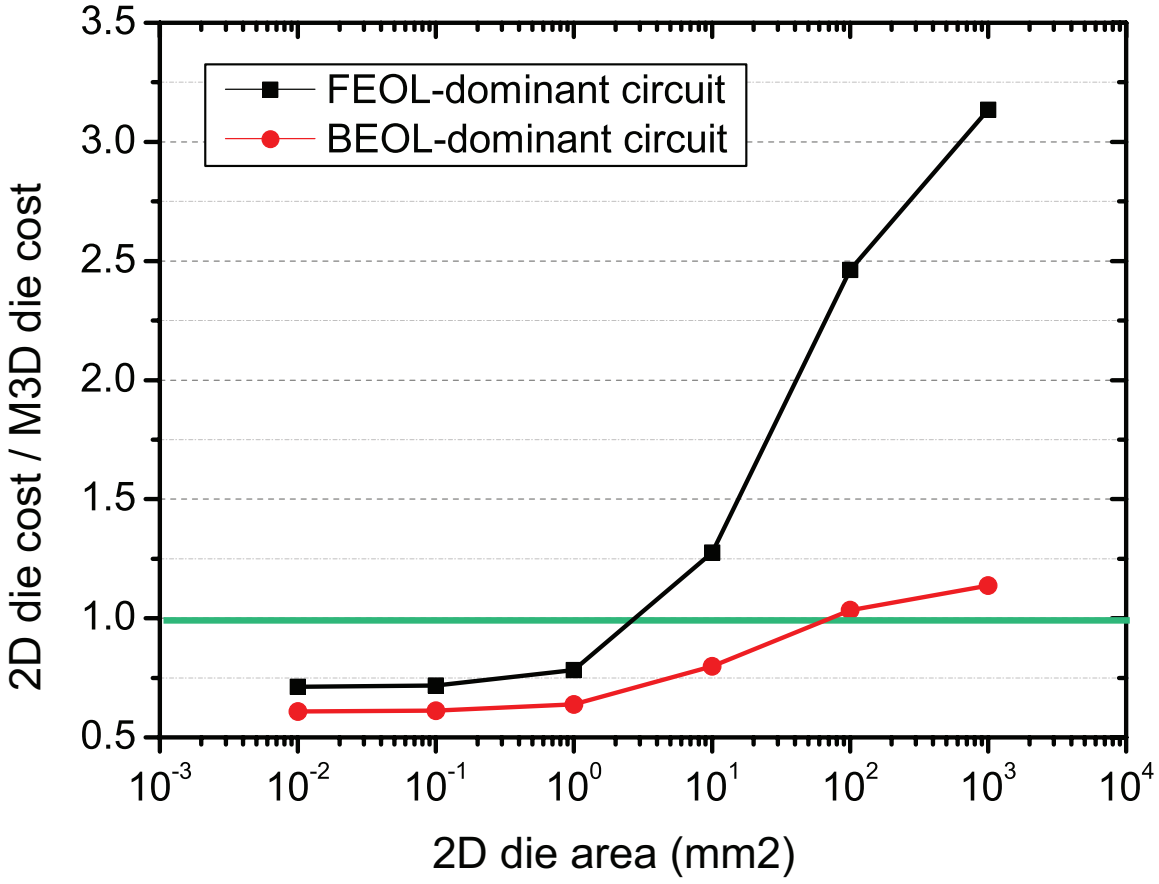


Figure 3.4: Die size impact on the die cost ratio between 2D and M3D. Two different circuit type (FEOL-dominant and BEOL-dominant) are investigated. The region above the green line indicates where the M3D die cost is cheaper than 2D die cost.

7nm bulk FinFET technology. We propose normalized wafer and die cost models based on the number of metal stacks and die area for 2D and M3D. In our PPC tradeoff study with the simple but self-contained cost models, both 2D and M3D designs are optimized in terms of the number of BEOL metal layers used for routing to obtain the best possible PPC values for the fair comparison. Also, a new CAD methodology for 2-tier G-M3D named Projected-2D Flow is developed. Projected-2D maximizes the placement and routing utilization of an M3D design by reducing its footprint by more than 50% compared with that of the 2D counterpart. Furthermore, this flow allows us to accurately compare RC parasitics of equivalent nets in both 2D and M3D designs since final netlists of these two design flavors are the same.

Based on the experiments with two widely different circuit types (BEOL-dominant vs. FEOL-dominant), it is confirmed that while M3D has indeed a great footprint saving, the PPC quality of M3D is actually worse than that of optimized 2D reference by 34% due to high M3D wafer cost. Our study also shows that, for the adoption of M3D technology at the 7nm era, M3D wafer yield needs to be higher than 90% of 2D wafer yield, and the 2-tier device manufacturing cost of M3D design needs to be limited by less than 33% of 2D device manufacturing cost, and lastly the die area should be large enough ($100mm^2$-scale) to have fruitful die cost reduction from huge M3D footprint saving. Lastly, and counter-intuitively, this study shows that FEOL-dominant type circuit has PPC benefits from M3D technology more and sooner than BEOL-dominant type circuit.

CHAPTER 4

PHYSICAL DESIGN SOLUTIONS TO TACKLE FEOL/BEOL DEGRADATION IN GATE-LEVEL MONOLITHIC 3D ICS

One of the most critical problems with M3D integration is the limited thermal budget for the top tier fabrication process. Once the bottom tier devices are implemented with the normal process, they suffer from additional thermal exposure during dopant activation step of the top tier ($T > 1000\,^{\circ}C$). In the meantime, integrating Copper (Cu) interconnects in the bottom tier implies that thermal budget for the top tier has to be under $450\,^{\circ}C$, because Cu diffuses away into the Low-K regions at such a high temperature. In order to preserve the device performance and the integrity of the back-end-of-line (BEOL) of the bottom tier, recent studies [49, 50] introduce molecular bonding and solid phase epitaxial regrowth (SPER) dopant activation process ($T \sim 450\,^{\circ}C$) for the top tier device manufacturing based on planar FDSOI device. In the industrial environment, however, implementing FinFET or nano-wire devices requires conformal in-situ doping in S/D region due to the 3D structure of the device, leading to high temperature annealing processes ($T > 1100\,^{\circ}C$). Therefore, the limited thermal budget for the top tier manufacturing is expected to bring serious performance loss on the device. Tungsten (W) interconnects in the bottom tier is an alternative to offer more thermal budget for the top tier manufacturing, but the high resistivity of W degrades the overall performance.

An earlier work [51] addresses inter-tier performance variations in M3D ICs. However, is was based on block-level M3D, where the design seriously under-utilizes MIVs and is thus not practical. The authors of [26] present a design methodology for 2-tier gate-level M3D, so-called Shrunk-2D flow. The drawback of this flow is that it requires the same RC parasitic despite shrinking geometry of layout objects. In the advanced node, parasitics are changing non-linear along with metal geometry. Therefore, it is possible to exaggerate

the parasitic value, leading to low quality M3D designs. Recently, M3D benefits have been studied for a predictive 7nm FinFET technology [42], but this work does not consider inter-tier variation caused by limited thermal budget.

In this research, we propose a new physical design solution for gate-level M3D that tackles the inter-tier performance variations caused by low temperature manufacturing. The key contributions of this work are as follows: (1) Using a 7nm bulk FinFET from a foundry-grade process design kit (PDK), we model the top tier device mobility degradation caused by the low thermal budget process, and show the impact on 2-tier gate-level full-chip M3D designs with Shrunk-2D flow. (2) We quantify the impact of both tungsten BEOL and cost-driven metal layer saving in the bottom tier on M3D design performance. (3) Using these transistor corners and interconnect models, we propose Derated-2D flow, where we do not alter the geometry in technology files, but only derate the RC parasitic corner. (4) We develop a tier partitioning algorithm, named Cell-Slack Sorting, that tries to assign timing critical elements into the bottom tier to address the top tier cell degradation. (5) We present a timing-driven MIV planning method and a post-route optimization flow that minimize the routing congestion and performance loss from the BEOL degradation. Experiments show that our design solutions allow only 3% performance degradation compared with 2D under harsh FEOL/BEOL variation settings.

4.1 FEOL/BEOL Variation Impact

Table 4.1 shows the nomenclatures used in this work. We choose Triple Data Encryption Standard Cipher (DES3) and Low-Density Parity Check Decoder (LDPC) from OpenCore benchmark suites [41] to cover different types of circuit. With regard to capacitance composition in Table 4.2, LDPC is a BEOL-dominant, and DES3 is a FEOL-dominant circuit. Figure 4.1 shows GDS layouts of their 2D implementation. All designs are implemented with foundry-grade 7nm bulk FinFET PDK. Using these two benchmarks, we factorize the impact of inter-tier variations caused by low temperature process on the performance of

full-chip 2-tier gate-level M3D design. The diameter of an MIV is assumed to be the width of top metal layer in the bottom tier (36nm for 5 metal, 24nm for 3 metal) with resistance of 16Ω and capacitance of $0.01fF$.

Table 4.1: Nomenclatures in this work.

TT	typical transistor corner (= no I_{on} degradation)
LT10p	10% I_{on} degradation in the top tier device
LT20p	20% I_{on} degradation in the top tier device
SVT	standard threshold voltage cell
LVT	low threshold voltage cell
Cu5	5 layers of copper BEOL used in the bottom tier
Cu3	3 layers of copper BEOL used in the bottom tier
W5	5 layers of tungsten BEOL used in the bottom tier
W3	3 layers of tungsten BEOL used in the bottom tier

Table 4.2: Our benchmark circuits, where the metrics are from 2D IC designs. All designs are implemented with a foundry-grade 7nm bulk FinFET technology.

	DES3	LDPC
Cell Count	44,978	59,297
Wire Cap : Pin Cap	28:72	64:36
Avg Net Length (*um*/net)	2.54	10.02
Avg Net Wire Cap (fF/net)	0.41	1.77
Avg Net Pin Cap (fF/net)	1.03	0.97
Circuit Type	FEOL-dominant	BEOL-dominant

4.1.1 Top Tier Device Degradation

The exact correlation between the low thermal budget process and the degree of top tier device degradation in the advanced node is not fully known. However, one of the main factors affected by low temperature process is expected to be the mobility of top tier devices. We illustrate the degree of top tier degradation for two scenarios, where 10% and 20% I_{on} decrease by mobility reduction. We refer to these degraded transistors as LT10p, LT20p corner, respectively. In order to evaluate the impact of mobility degradation on a device, we measure I_{on}, I_{off} of the device by sweeping mobility parameter. We use 7nm bulk FinFET Standard V_{TH} (SVT), and Low V_{TH} (LVT) compact model from a foundry-grade PDK with the nominal supply voltage 0.65V. From simulation results, we observe that 18.2%

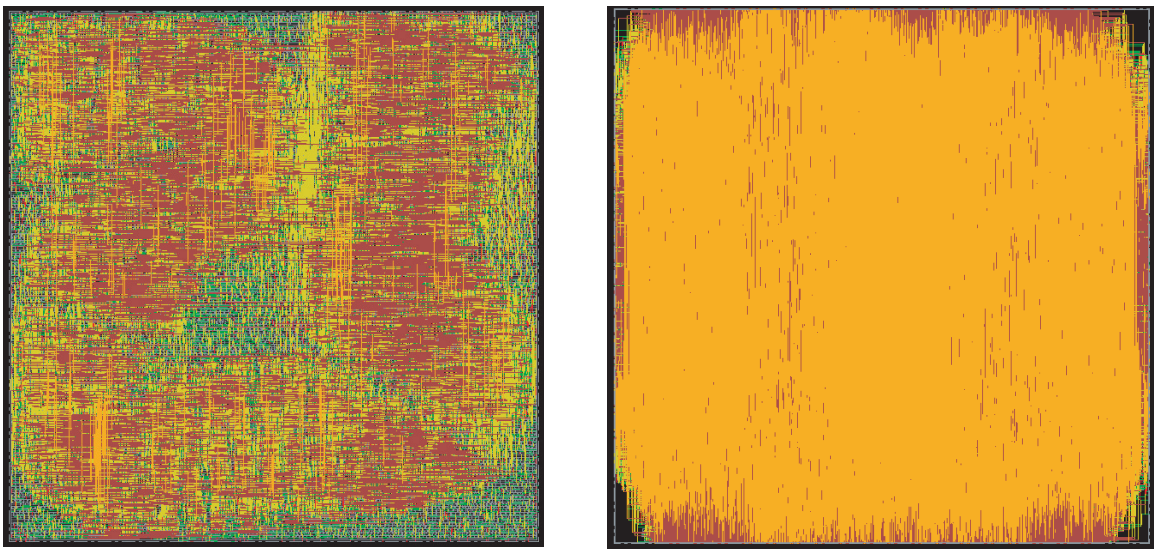


Figure 4.1: GDS layouts of 2D designs of our benchmark.

and 33.4% mobility degradation cause 10% and 20% I_{on} decrease, and 18.5% and 34.0% I_{off} reduction, respectively.

To analyze cell-level performance degradation, we characterize standard cell libraries for LT10p and LT20p corners with Cu local interconnect using Virtuoso Liberate. Using these cell models, we measure output slew and gate delay of cells assuming 10ps input slew and FO3 inverters with 300 Contacted Poly-Pitch (CPP $= 42nm$)-length M2 wire loading using Synopsys PrimeTime. Table 4.3 shows that LT10p and LT20p corners result in 10.0%, 22.7% of cell performance degradation, respectively.

Table 4.3: Impact of mobility degradation on cell performance. We show the average output slew and delay in (ps) among INVx1, ND2x1, XNR2x1, AOI22x1, and DFF Clk-Q. Copper local interconnects are used.

	TT, Cu	LT10p, Cu	LT20p, Cu
Avg. SVT output slew	22.72 (1.00)	25.16 (1.11)	28.37 (1.25)
Avg. SVT cell delay	86.29 (1.00)	94.61 (1.10)	105.05 (1.22)
Avg. LVT output slew	16.62 (1.00)	18.27 (1.10)	20.32 (1.22)
Avg. LVT cell delay	57.28 (1.00)	62.90 (1.10)	69.81 (1.22)

Starting from ideal scenario where there is no degradation on top tier devices and equiv-

alent 5 metal layers of Cu BEOL in both tiers, Figure 4.2 shows the impact of top tier device degradation on the maximum performance of full-chip 2-tier gate-level M3D design. We use Shrunk-2D flow [26, 43] based on foundry-grade 7nm bulk FinFET PDK. Since Shrunk-2D flow does not handle the inter-tier variation, we modify only the last tier-by-tier routing stage to consider the inter-tier performance variation as described in Section **??**. Assuming 20% I_{on} reduction on the top tier, the performance of DES3, and LDPC is degraded by 21%, and 10% respectively. DES3 is a FEOL-dominant circuit, and 99% of the longest path delay consists of cell delays. Therefore, the performance of DES3 is sensitive to the FEOL degradation. On the other hand, LDPC has net delays of 21% out of the longest path delay, so the impact of cell delay increase is less than that of DES3. The simulation result identifies top tier device degradation as one of the critical obstacles to meet the timing of M3D design.

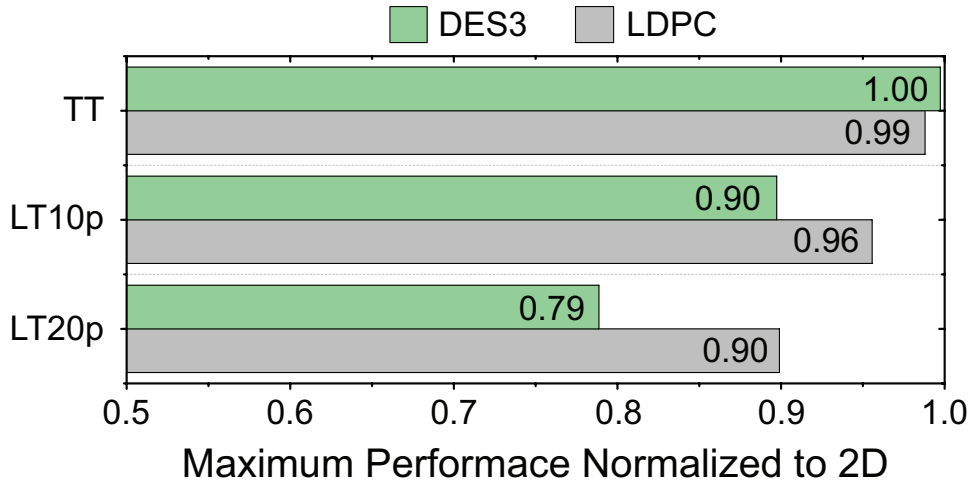


Figure 4.2: Impact of top tier device degradation on full-chip 2-tier M3D performance. We use 5 layers of Cu BEOL in both tiers. DES, our FEOL-dominant circuit, is more sensitive to the degradation.

4.1.2 Bottom Tier Interconnect Degradation

In order to provide more thermal budget for the top tier manufacturing, integrating W BEOL in the bottom tier is an alternative. To quantify the impact of W interconnect, we

first modify the process file (ICT) from foundry-grade 7nm PDK assuming W conducting layers and TiN liners with same geometry as Cu BEOL, and generate QRC technology file (TCH) using Cadence Techgen. Next, we characterize standard cell libraries based on W local interconnect using Virtuoso Liberate. Since wirelength of local interconnect is very short in the cell layout, we observe only 2% slew degradation and 1% output delay increase in both SVT and LVT cells. Based on these interconnect and cell models, Figure 4.3 shows the impact of tungsten interconnect in the bottom tier on the maximum performance of full-chip 2-tier M3D designs. We assume no device degradation on the top tier. We observe that M2, M3 layers have 2.20 times as high as Cu resistance (ohm/um), and 2.46 times for M4, M5 layers. When we compare the maximum performance under 5 layers of Cu BEOL (Cu5) with the result under 5 layers of W BEOL (W5) case, LDPC has 21% performance degradation while the performance of DES3 is decreased by only 2%. This is because net delays of BEOL-dominant LDPC are significantly increased due to the highly resistive W interconnect. DES3 has minor performance degradation since most of the path timing consists of cell delays.

Another interesting perspective on the bottom tier BEOL is to reduce the number of metal layers. BEOL cost increases significantly from N28 to N7 nodes due to dimensional scaling and multiple patterning processes [48]. Therefore, reducing metal stack must be taken into account to make M3D cost-effective. We consider 2 metal layers saving from the bottom tier in Figure 4.3. When we compare Cu3 result with Cu5 case, both DES3 and LDPC have significant performance loss. Reduced routing resources cause huge routing congestion in the bottom tier, resulting in 14% capacitance increase and 17% resistance increase on average. Under the worst scenario where we use W BEOL and reduce 2 metal layers from the bottom tier, 16% and 39% of maximum performance is degraded in DES3 and LDPC, respectively.

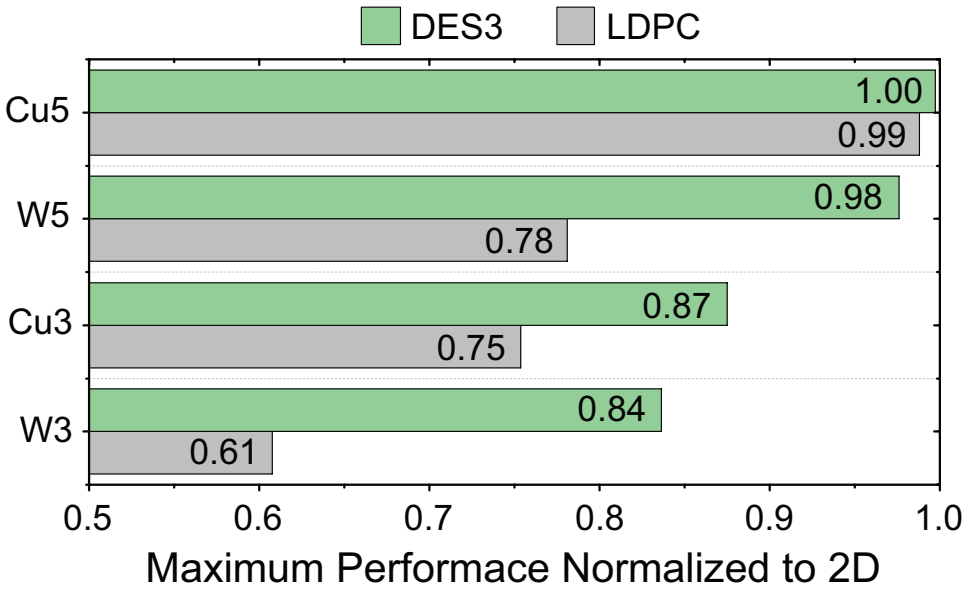


Figure 4.3: Full-chip impact of tungsten BEOL and metal layer saving in the bottom tier. LDPC, our BEOL-dominant circuit, is more sensitive to the changes.

4.2 Physical Design Solutions

To tackle the inter-tier performance variations caused by top tier low temperature manufacturing, we propose a new full-chip M3D physical design flow named Derated-2D. Four CAD methodologies are proposed in Derated-2D flow as follows: (1) We do not modify the layout objects but complete the 2D IC design with derated RC parasitic corner, named Derated-2D design. Then we project the placement result of Derated-2D design into the final footprint of M3D design. (2) For the low-temperature process-aware tier partitioning, we use cell-slack as a metric for the timing criticality, and assign the timing critical elements into the bottom tier to address top tier cell degradation. (3) Timing-driven MIV planning deals with resistive W interconnect and reduced metal stack in the bottom tier. (4) A post-route optimization flow compensates the performance degradation under various FEOL/BEOL degradation settings at a minimum energy overhead. Overall design methodologies for Derated-2D flow is shown in Figure 4.4.

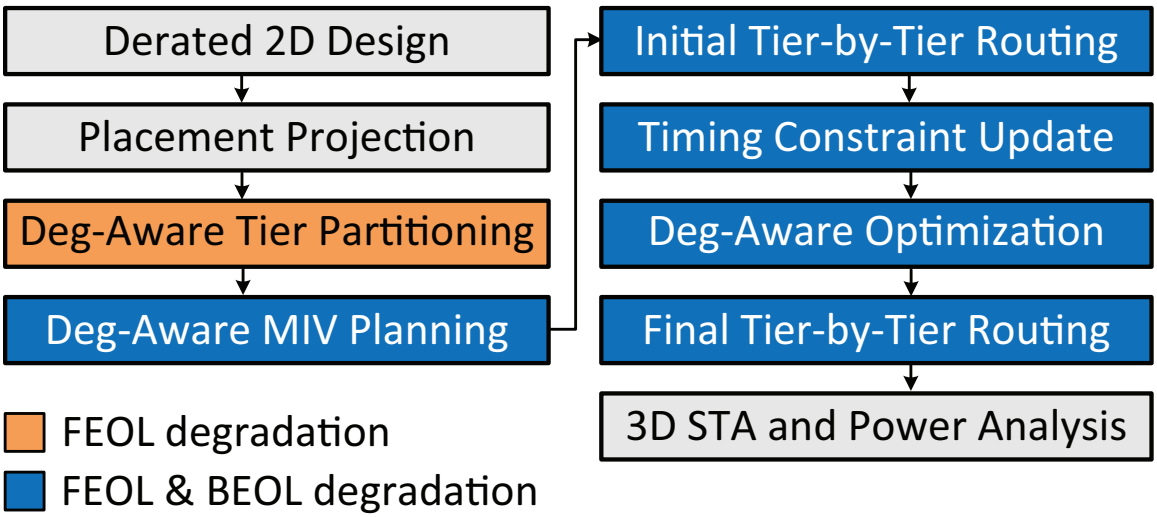


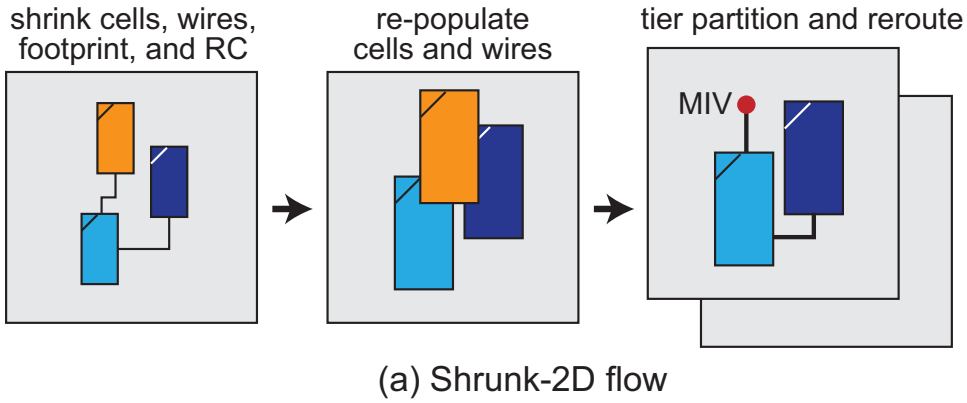
Figure 4.4: Derated-2D, our FEOL/BEOL degradation-aware physical design flow for gate-level M3D. Our tier partitioning step tackles FEOL degradation, while the subsequent steps address both FEOL and BEOL degradation.

4.2.1 Derated-2D Design and Projection

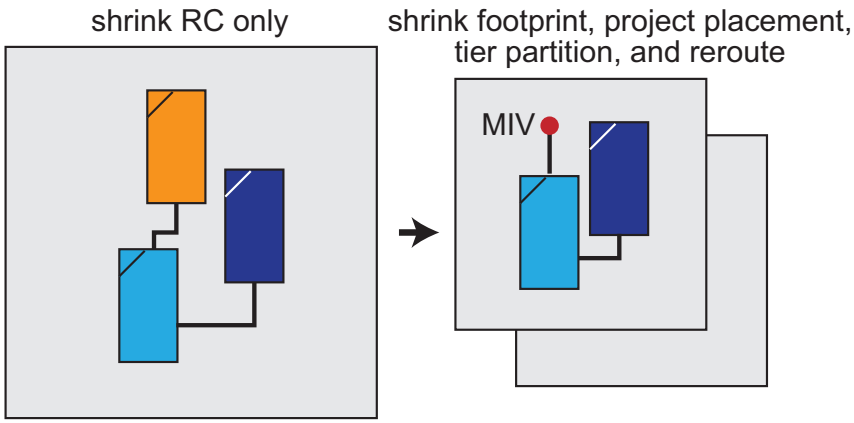
Unlike Shrunk-2D flow [26] that requires shrinking of layout objects and RC parasitic scaling, Derated-2D uses original layout objects. However, Derated-2D is also possible to have overestimated wire load and redundant buffers, unless we consider the wirelength saving from reduced footprint of M3D design. Assuming no silicon area overhead, 2-tier M3D design has half footprint of that of 2D. In order to optimize Derated-2D design with same RC parasitics of M3D design, we first create an RC corner which is derated by $1/\sqrt{2}$ for total R and total C while not scaling coupling capacitance due to the same routing pitch in Derated-2D and M3D. Then we project the whole placement result of Derated-2D design into the footprint of final M3D design. Since every manhattan distance between each macro is scaled by $1/\sqrt{2}$ as a result of placement projection, RC parasitic of Derated-2D design is expected to be the same as that of M3D design. Table 4.4 and Figure 4.5 show comparison between our Derated-2D flow and state-of-the-art Shrunk-2D flow.

Table 4.4: Comparison between our Derated-2D flow and state-of-the-art Shrunk-2D flow [26].

	Derated-2D	Shrunk-2D
Shrink chip footprint?	No	Yes
Shrink cell layout?	No	Yes
Shrink metal dimension?	No	Yes
Scale unit-length RC parasitics?	Yes	Yes
Consider FEOL degradation?	Yes	No
Consider BEOL degradation?	Yes	No
Bottom tier cells use top tier metal?	Yes	No
Post-route optimization supported?	Yes	No



(a) Shrunk-2D flow



(b) Derated-2D flow

Figure 4.5: Illustration of Shrunk-2D [26] and Derated-2D flow.

4.2.2 Tier Partitioning and MIV Planning

Cell slack is a metric to measure how long each cell may delay without compromising the timing of paths propagating through the cell. We extract the slack value of timing

constrained cells on the Derated-2D design using Synopsys PrimeTime, and use it as a metric to represent timing criticality of a cell. For the clock network cells, we keep them to the bottom tier. The simplest partitioning scheme using cell slack is to sort them in decreasing order, and if the slack of a cell is less than median, then to place it on the bottom tier. In most cases, however, these timing critical cells are usually placed close to each other, resulting in local area skew between each tier. With unbalanced area skew, cell slack sorting does not guarantee minimum performance degradation since the original location of a cell that is already optimized at Derated-2D design must be changed during placement legalization. Therefore, we divide the design footprint by small size partitioning bin in the regular fashion, and sort cells within each bin by the slack value in decreasing order to meet the local area balance.

For the MIV planning, Shrunk-2D flow introduces CAD methodology to manipulate commercial engine built for 2D ICs [43]. Our timing-driven MIV planning in Derated-2D flow also uses the basic idea of MIV planning scheme in Shrunk-2D flow but the differences are as follows: (1) In the same way that we create a 3D LEF to define two macro flavors for each standard cell - one for each tier, we create a 3D LIB that defines two timing flavors to consider inter-tier device performance variation. Thus, timing model for each cell in 3D space is mapped to its appropriate transistor corner. We import the 3D LIB into commercial router (Candence Innovus) to create delay corner for timing-driven routing. (2) Since each tier is possible to have different routing material and number of metal layers, we create a process file (ICT) for full 3D metal stack and generate 3D TCH file using Cadence Techgen. This 3D TCH file contains the RC parasitic information for every routing layers in M3D design. Then, we create a parasitic corner with this 3D TCH file in the commercial route. With timing constraint same as 2D design, we do timing-driven routing to insert MIVs. Using full 3D metal stack makes it possible to share the routing resources from all tiers. If we reduce the metal stack on the bottom tier, then the router uses top tier metal layers to route bottom tier nets in order to minimize routing congestion. If W BEOL is used in

the bottom tier, the tool tries to use low resistive Cu BEOL on the top tier to minimize timing degradation. Figure 4.6 shows the differences of MIV planning scheme between Shrunk-2D and Derated-2D flow.

4.2.3 Post-Route Optimization and Routing

Since the initial Derated-2D design only involves normal transistor corner and Cu BEOL, it is clear that there exists limitation for timing closure under inter-tier variations in M3D design. Since we create delay and parasitic corners at timing-driven MIV planning stage, it is also possible to use a post-route optimization flow to update initial Derated-2D design for timing closure at a mimimum energy overhead. We update the timing constraint for a post-route optimization in consideration of cell legalization during tier-by-tier routing. Then, we change the size of macros in 3D LEF into the size of placement site, which is the smallest dimension that a macro can have. By using unit size 3D macros, we remove the placement overlap and again minimize the cell legalization during post-route optimization. We keep the location of initial top tier cells, and allow the commercial tool to optimize bottom tier cells by resizing and VT swapping for timing closure. The reason why we do not play with the top tier cells is that the MIV routing blockages are initially fixed under the placement result of top tier cells. After post-route optimization, we proceed final tier-by-tier routing to create separate GDS for each tier.

Once the MIV locations are determined by MIV planning, we create a DEF file for each tier containing the location of MIVs as primary I/O. Using original macro LEF, we repopulate the cell size and legalize the placement overlap. We route them initially with appropriate LIB (TT,LT10p,LT20p) and TCH (Cu,W) to the specific FEOL/BEOL degradation scenario, and create the timing context of each tier to optimize the routing quality. After routing under the timing context, we extract the parasitic, and proceed to 3D timing and power analysis using Synopsys PrimeTime.

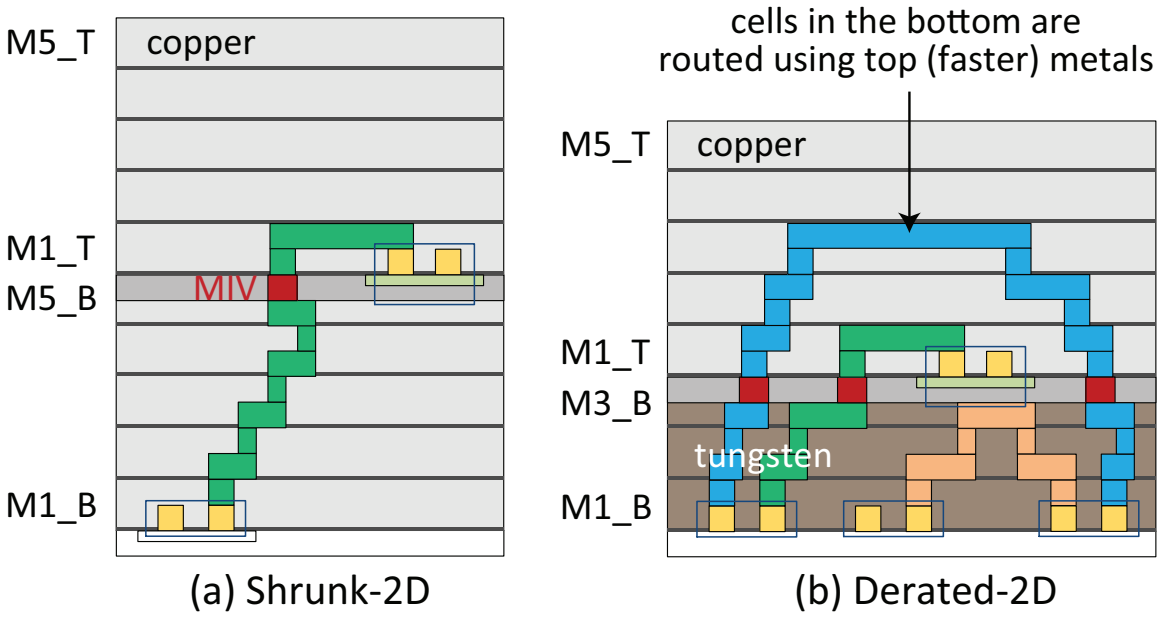


Figure 4.6: Metal stack comparison. (a) Shrunk-2D [26] with 5 Cu metal layers in both tiers, (b) Derated-2D flow with 5 layers of Cu in the top, and 3 tungsten in the bottom. Top cells contain MIV routing obstacle underneath.

4.3 Experimental Results

4.3.1 Impact of Tier Partitioning

Figure 4.7 shows the impact of cell-slack sorting tier partitioning on the design performance compared with Fiduccia-Mattheyes (FM) min-cut partitioning algorithm [26]. To be an equal comparison, we use Derated-2D design for both of the partitioning algorithms, and assume 5 layers of Cu BEOL in both tiers. Even under 20% I_{ON} degradation on the top tier, cell-slack sorting partitioning allows only 5% of performance degradation in both benchmarks. Table 4.5 shows detailed statistics of M3D designs from different partitioning algorithms. Min-cut partitioning tries to minimize the connections between each tier inside the partitioning bin. Therefore, 2D nets on each tier get longer and congested, leading to further longer 3D nets. However, cell slack sorting partitioning uses as many MIVs as necessary in order to assign the timing critical cells to the bottom tier. While minimizing the impact of top tier cell delay increase, these many and short 3D connections also effectively

reduce net delay, resulting in significant timing saving. The incremental gain update makes FM min-cut heuristic run in $O(C)$, where C is the number of cells. Cell-slack sorting runs in $O(C \log C)$ by sorting algorithm.

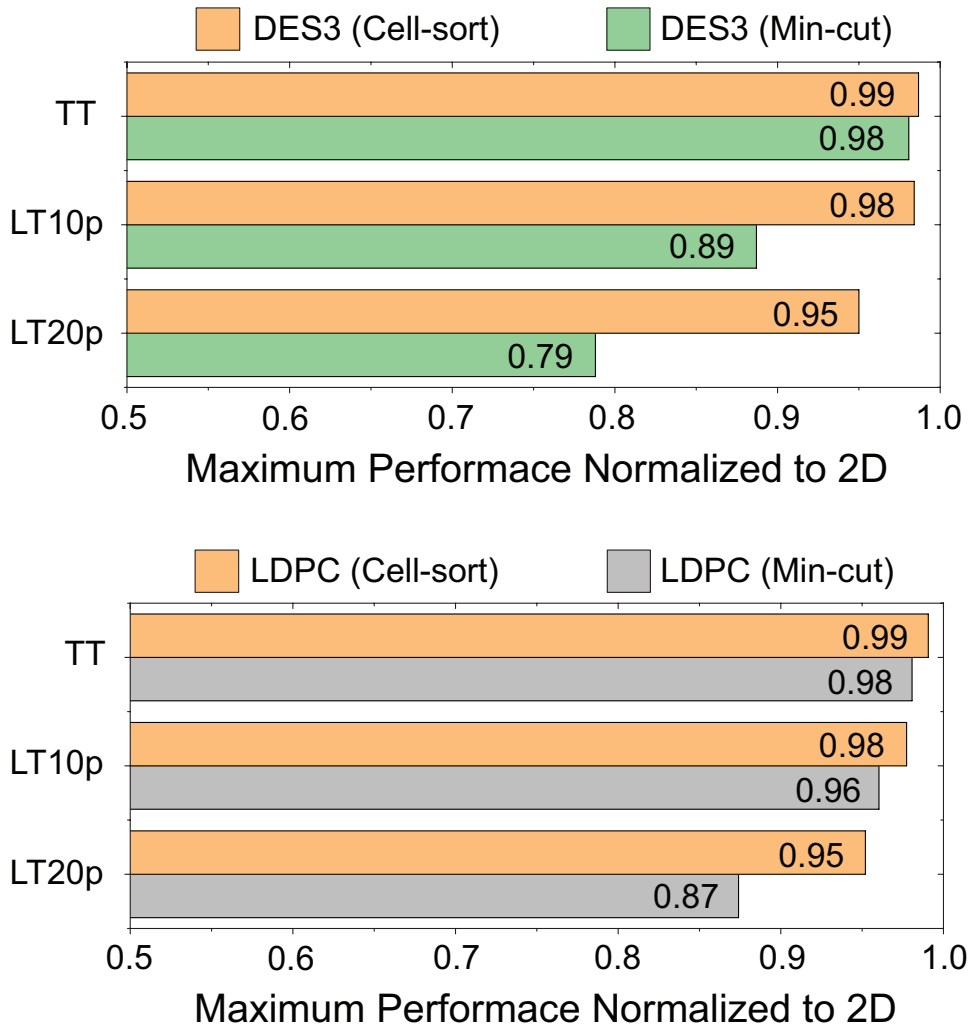


Figure 4.7: Tier partitioning impact on performance under FEOL degradation. Our cell sorting-based method withstands the degradation better than min-cut for both circuits.

4.3.2 Impact of MIV Planning

Based on cell slack sorting tier partitioning, Figure 4.8 shows the impact of our timing-driven MIV planning compared with Shrunk-2D flow. Under no top tier device degradation, W BEOL and 2 metal layer reduction in the bottom tier leads to 23% and 36% performance

Table 4.5: Comparison between cell-slack sorting vs. min-cut tier partitioning. We use LT20p transistor corner in the top tier, and 5 layers of Cu BEOL in both tiers.

	LDPC		DES3	
	min-cut	cell-sort	min-cut	cell-sort
Cell Count	57451		44805	
Net Count	59696		45036	
2D Net (top tier) Count	23171	16284	16677	10289
2D Net (bot tier) Count	24118	21718	23063	13701
3D Net Count	12407	21694	5296	21046
MIV Count	25958	37189	6772	25500
Avg. MIV# of 3D Net	2.09	1.71	1.28	1.21
Avg. WL of 2D Net (um/net)	3.29	2.83	2.18	1.69
Avg. R of 2D Net (ohm/net)	416.81	372.45	342.47	268.52
Avg. WL of 3D Net (um/net)	23.42	14.72	5.60	3.91
Avg. R of 3D Net (ohm/net)	2692.18	1743.30	867.53	650.95
Target Clock (ns)	1.0	1.0	0.5	0.5
WNS (s)	-0.16	-0.07	-0.18	-0.06
TNS (s)	-68.59	-2.86	-52.46	-3.58
TPS (s)	19.85	90.87	2605.71	2618.53
Runtime (sec)	24	111	12	44

degradation in DES3 and LDPC with MIV planning in Shrunk-2D flow. Our timing-driven MIV planing, however, allows only 13% and 20% of the performance degradation in DES3 and LDPC respectively. In Table 4.6, we analyze net distribution and parasitics of LDPC design. Since 2D nets are possible to become 3D nets with our timing-driven MIV planning to close the timing, nets in the resistive bottom tier are routed by Cu BEOL in the top tier. Also, sharing routing resources between each tier decreases average net length and RC parasitics on the bottom tier and balances the routing congestion caused by metal layer reduction. Therefore, net delay degradation caused by W BEOL and routing congestion on the bottom tier are minimized, resulting in performance saving. In addition, top tier device degradation has a minor impact on design performance when bottom nets are in the worst scenario as a result of cell-slack sorting tier partitioning.

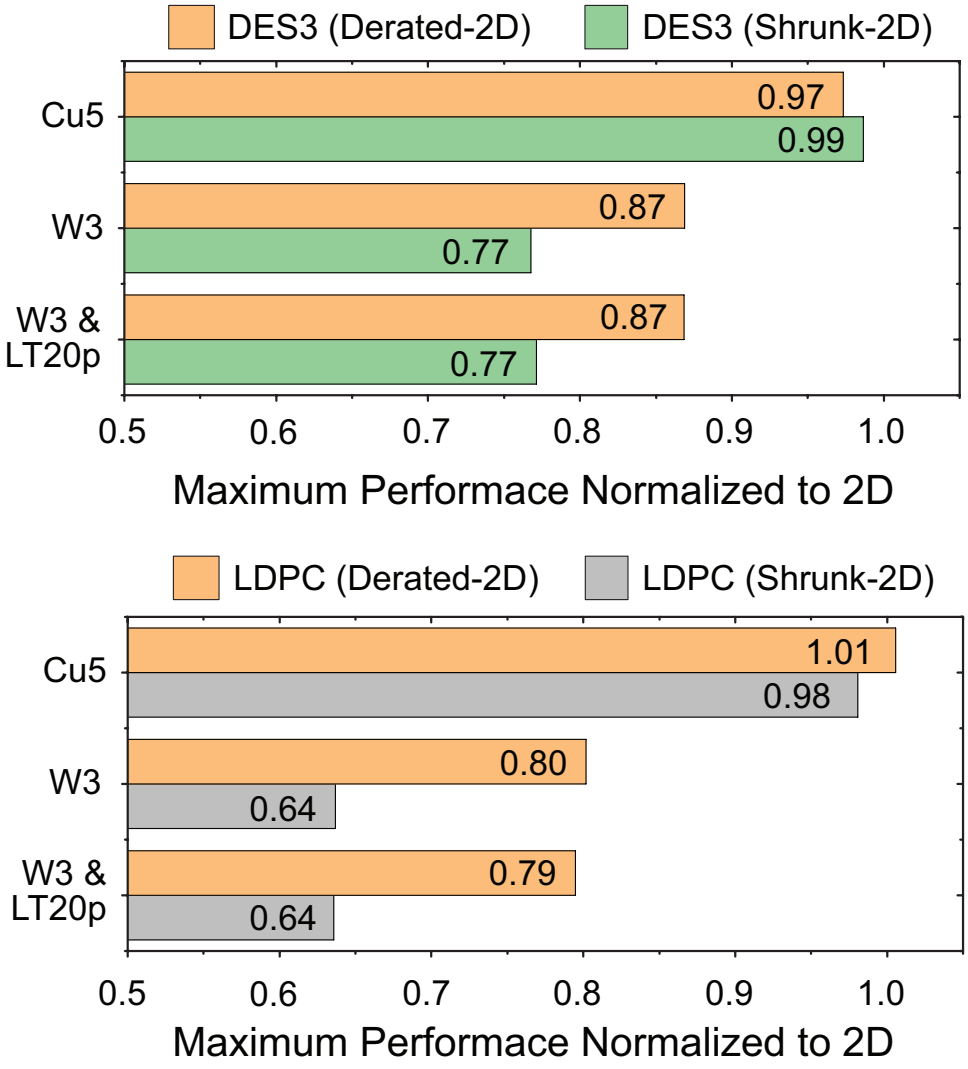


Figure 4.8: Impact of MIV planning in Derated-2D vs. Shrunk-2D [26]. Our Derated-2D withstands the FEOL and BEOL degradation better than Shrunk-2D.

Table 4.6: Comparison between MIV planning in Shrunk-2D [26] vs. our Derated-2D. We assume no FEOL degradation and use 3 tungsten BEOL layers in the bottom tier in LDPC benchmark. Derated-2D encourages more routing in the top tier (= faster Cu BEOL).

	metric	Shrunk-2D	Derated-2D
net stats	top placed, top routed	17,432	17,410
	top placed, top/bot routed	0	22
	bot placed, bot routed	22,280	19,072
	bot placed, top/bot routed	0	3,208
	top/bot placed, top/bot routed	19,984	19,984
top tier	Avg. Net Length (um/net)	5.40	**6.85**
	Avg. Net Cap (ff/net)	2.70	**2.92**
	Avg. Net Wire Cap (ff/net)	0.92	**1.24**
	Avg. Net Res (Ohm/net)	601.81	**758.12**
bot tier	Avg. Net Length (um/net)	**3.50**	2.64
	Avg. Net Cap (ff/net)	**2.62**	2.45
	Avg. Net Wire Cap (ff/net)	**0.78**	0.52
	Avg. Net Res (Ohm/net)	**1192.32**	916.06
	Avg. MIV# per 3D net	2.1	1.6
	Max. Performance (GHz)	**0.68**	**0.75**
	Power-Delay Product (pJ)	32.59	32.22

Table 4.7: Performance and power-delay product (= energy) comparison under various FEOL and BEOL degradation settings. Our Derated-2D consistently outperforms Shrunk-2D [26] in terms of both performance and energy, even in the worst-case scenario (20% slow device, 3 layers of tungsten routing). Our post-route optimizer further improves performance at the expense of energy increase.

FEOL/BEOL setting		Maximum performance normalized to 2D			Post-route Optimization impact on TNS		Power-Delay Product normalized to 2D		
top tier	bot tier	Shrunk-2D	Derated-2D	D2D+PostOpt	Derated-2D	D2D+PostOpt	Shrunk-2D	Derated-2D	D2D+PostOpt
LDPC									
TT, Cu5	TT, Cu5	0.98	1.01	1.01	-0.01	-0.01	0.84	0.78	0.78
TT, Cu5	TT, Cu3	0.78	0.91	0.99	-13.06	-0.27	0.92	0.84	0.85
LT10p, Cu5	TT, Cu3	0.77	0.89	0.98	-16.05	-0.33	0.92	0.84	0.85
LT20p, Cu5	TT, Cu3	0.75	0.84	0.98	-38.12	-0.28	0.92	0.84	0.86
TT, Cu5	TT, W3	0.61	0.80	0.98	-137.90	-0.12	0.93	0.85	0.90
LT10p, Cu5	TT, W3	0.60	0.79	0.98	-159.11	-0.33	0.93	0.85	0.90
LT20p, Cu5	**TT, W3**	**0.58**	**0.79**	**0.97**	-186.12	-0.19	**0.93**	**0.84**	**0.92**
DES3									
TT, Cu5	TT, Cu5	1.00	0.97	1.03	-0.82	-1.08	0.98	0.98	0.98
TT, Cu5	TT, Cu3	0.87	0.90	1.02	-4.16	-1.78	1.01	0.99	1.00
LT10p, Cu5	TT, Cu3	0.86	0.90	1.03	-4.42	-1.43	1.01	0.99	1.00
LT20p, Cu5	TT, Cu3	0.79	0.90	1.03	-7.15	-2.32	1.01	0.99	1.01
TT, Cu5	TT, W3	0.84	0.87	1.01	-8.97	-2.77	1.02	1.00	1.01
LT10p, Cu5	TT, W3	0.82	0.87	1.03	-8.80	-2.44	1.02	1.00	1.01
LT20p, Cu5	**TT, W3**	**0.78**	**0.87**	**1.02**	-11.33	-1.96	**1.02**	**1.00**	**1.01**

4.3.3 Comparison with Shrunk-2D Flow

Based on a foundry-grade 7nm FinFET PDK, we compare the design results of our Derated-2D flow with results of Shrunk-2D flow under all inter-tier variation scenarios in Table 4.7. Under the worst scenario, where there is 20% I_{on} reduction in the top tier while saving 2 metal layers of W BEOL in the bottom tier, Derated-2D result of LDPC achieves 36% of performance improvement, and 10% of energy saving compared with Shrunk-2D result without post-route optimization.

When we compare the design results between LDPC and DES3, we observe that energy saving from M3D depends on the circuit type. LDPC, which is a BEOL-dominant circuit, has energy saving of 22% in ideal scenario, and still has 16% saving under the worst scenario without post-route optimization. This is because major source of energy saving from M3D design is wirelength reduction. In 2-tier M3D design with Derated-2D flow, expected maximum total wirelength saving is 29.3% considering 50% footprint saving from M3D design. Although the routing congestion in each tier is possible to degrade the wirelength saving depending on the circuit, this huge wirelength saving leads to around 30% of wire capacitance saving, and if the design is BEOL-dominant such as LDPC that 64% of total capacitance is wire capacitance, total capacitance saving becomes 22%. This capacitance saving is directly converted into switching power saving of 22%. However, since the total power consists of switching power, internal power, and leakage, the final power saving become 16%. In the case of DES3, wire capacitance is only 28% of total capacitance. Therefore, switching power saving from 30% of wirelength reduction in M3D is degraded into only 8%, and the final power saving degraded by internal power ratio become 2%.

We also tabulate the impact of a post-route optimization flow on timing and power-delay product. Under any scenarios, the post-route optimization makes it possible to restore the performance degradation of M3D design up to minimum 97% of 2D performance. Under the worst scenario where 20% I_{on} degradation on the top tier and W BEOL with 2 metal layer saving in the bottom tier, we recover the M3D performance of LDPC from 79% to

97% of 2D performance at the expense of 8% of energy. In case of DES3, we observe that although FEOL-dominant circuit has less energy saving, since it is less affected by resistive W interconnect and bottom routing congestion than BEOL-dominant circuit, it requires only 1% of 2D energy to restore the performance degradation.

4.4 Conclusion

In this research we proposed CAD methodologies for gate-level monolithic 3D ICs (M3D) that tackle the FEOL/BEOL inter-tier variations caused by low temperature manufacturing. To address the top tier device degradation, we presented a cell-slack sorting-based tier partitioning algorithm that assigns timing critical elements into the bottom tier. To deal with the BEOL impact, we developed a timing-driven MIV planning flow and a post-route optimization flow to compensate for the reduced routing layers and increased resistance of tungsten interconnect. Experiments along with 7nm bulk FinFET from a foundry-grade PDK demonstrated the effectiveness of our approach.

CHAPTER 5
COMPACT-2D: A PHYSICAL DESIGN METHODOLOGY TO BUILD COMMERCIAL-QUALITY FACE-TO-FACE 3D ICS

5.1 Gate-level Face-to-Face 3D Integration

Face-to-face (F2F) bonding technology involves 3D integration of two pre-fabricated dies in a face-to-face fashion. In F2F bonding, electrical connections between the dies are made by F2F vias, and the minimum pitch of these F2F vias defines the density of 3D interconnections. As 2D interconnects become denser along with logic device scaling, it calls for a tighter 3D interconnect pitch to improve the functional density and power-performance-area benefit of F2F-bonded 3D ICs. To enable smaller F2F via pitches, R&D has focused on enhancing the bonding precision of F2F integration lately. Among several notable achievements, hybrid wafer-to-wafer (W2W) bonding technology has emerged as a promising solution.

Hybrid W2W bonding is a wafer-level integration technology that enables direct metal-to-metal (damascene-pad) and dielectric-to-dielectric bonding between the back-end-of-lines (BEOLs) of pre-fabricated wafers [12, 13]. Thanks to the high precision of wafer-level integration, the minimum pitch is projected down to $0.8\mu m$ in the near future [16, 17]. This allows designers to utilize fine-grained and silicon-space overhead-free 3D interconnections in F2F-bonded 3D ICs.

Various studies have shown the benefits of F2F-bonded 3D ICs. Using a $5\mu m$ F2F pitch, [52] demonstrated a test chip that achieves a high memory bandwidth ($63.8GB/s$) in core-memory stacking architecture at $4W$ power consumption. [53] adopted F2F bonding technology for the heterogeneous integration of MEMS and SoCs, and reported 30% form factor savings. However, all these benefits are still based on a large F2F via pitch. To fully

benefit from the advanced F2F integration technology, new design and CAD solutions are required. In this research, we present a physical design methodology named Compact-2D (C2D) to build high-density and commercial-quality F2F-bonded 3D ICs.

5.2 Motivation

Previously, Shrunk-2D flow [26] is used to build F2F full-chip designs. The overall flow and issues of Shrunk-2D presented in the previous chapters are again emphasized in detail here due to its criticality. S2D requires shrinking of standard cells and interconnects by 50% to fit into the footprint of a two-tier F2F design with no silicon-area overhead. Then, the shrunk layout objects are used to implement the Shrunk-2D design, where the (X,Y) locations of cells are optimized with the same half perimeter wirelength (HPWL) as that of the target F2F design, assuming that the Z dimension is so small and thus negligible. To decide the Z location of each cell, tier partitioning is subsequently performed. Then, F2F via planning decides the actual F2F via locations based on the (X,Y,Z) placement solution. Although S2D shows how to use commercial 2D P&R engines to design F2F-bonded 3C ICs, it introduces the following new issues, especially in the advanced technology nodes.

- To handle shrunk geometries, S2D requires place/route (P&R) engines and design rule checkers that target one node smaller technology, which is both challenging and costly.[1]
- The shrunk dimension of interconnects leads to inaccuracy in RC parasitics of the S2D design unless the parasitic database is rebuilt for the shrunk geometries.[2]
- Tier partitioning in S2D ignores the fact that any inter-tier 3D route requires the *full* metal stacks for *both* tiers in F2F designs. Nevertheless, S2D does not support any

[1] Our conversations with S2D flow users at industry design houses revealed an exponential increase in design rule violations at the 7nm node. S2D suggests that designers ignore these errors. However, they reported that an excessive number of violations may cause commercial engines to terminate abruptly or produce low-quality layouts.

[2] Unless the resistivity and thickness of an interconnect are modified, the unit length resistance of a wire segment is not the same in S2D and F2F designs because the width of the interconnect is shrunk by 29.3%. Similarly, the shrunk width and spacing of the interconnect lead to inaccurate capacitance values in S2D designs.

optimization after tier partitioning. Therefore, it is prone to timing failure caused by inter-tier 3D routing overheads.

The physical design of monolithic 3D ICs [54, 29, 55] resembles that of F2F-bonded 3D ICs because the inter-tier vias are negligibly small. This offers similar freedom in constructing a (X,Y,Z) placement solution for both monolithic 3D and F2F designs. However, a notable difference lies in how inter-tier routing is done: in monolithic 3D ICs, only a single stack of BEOL is used, whereas both stacks are required in F2F-bonded 3D ICs. This motivates us to address the inter-tier 3D routing overheads efficiently (in both timing and power) for commercial-quality F2F designs. Thus, existing works on monolithic 3D ICs cannot be easily migrated to handle F2F designs.

5.3 Design Methodology

This section presents our design methodology named Compact-2D (C2D) flow to build commercial-quality F2F-bonded 3D ICs. C2D flow finds the (X,Y) placement solution of a F2F design using the original geometries of standard cells and interconnects. It also introduces an optimization capability to take the inter-tier 3D routing overheads into account correctly. The overall design methodology is shown in Figure 5.1.

Table 5.1: Terminologies in our Compact-2D (C2D) flow.

Compact-2D Design	An initial 2D design with unit length RC scaled by a factor of 0.707
Memory Expansion	Expanding the pin locations and memory macro boundaries by a facror of 1.414
Placement Contraction	Linearly contracting the placement solution of a Compact-2D design by a factor of 0.707
Compact F2F Via Planning	Performing timing, power, and F2F via location co-optimization to address inter-tier routing overhead
Incremental Routing	Recycling the routing result from Compact F2F Via Planning for the final GDSII generation

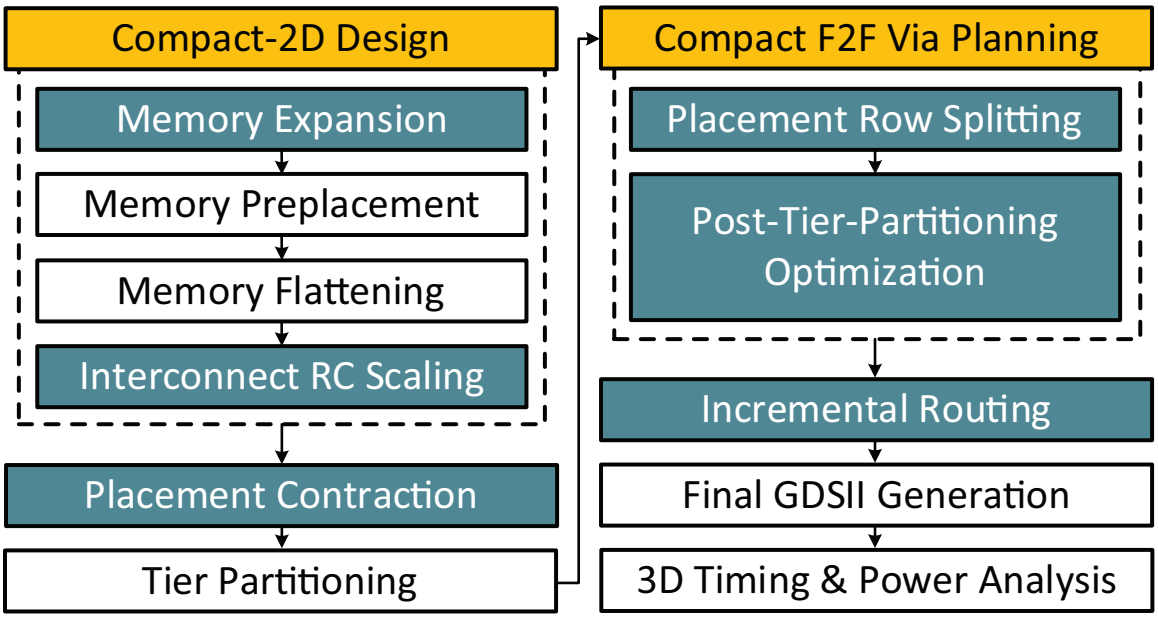


Figure 5.1: Our Compact-2D (C2D) flow. In color are the key steps proposed in this research to build commercial-quality F2F-bonded 3D ICs using 2D IC implementation tools.

5.3.1 Compact-2D Design

A Compact-2D design is a pseudo-3D design in C2D flow to find the optimal (X,Y) locations of standard cells in a target F2F design. The floorplan of the Compact-2D design is two times as large as the final 3D footprint in the same aspect ratio to accommodate all the synthesized gates in the two-tier F2F design with their original geometries. However, the HPWL of a net in the F2F design is 29.3% shorter than the corresponding net in the Compact-2D design when both are projected on the X-Y plane. To match the electrical length despite the difference in geometrical length, Figure 5.2 illustrates the need for interconnect RC scaling in the Compact-2D design. By Scaling the unit RC per length by a factor of 0.707, we avoid the redundant buffer insertions caused by increased geometrical length in the Compact-2D design while still using the original geometries of standard cells and interconnects. Then, we perform all the required implementation steps of the conventional 2D ICs in the Compact-2D design.

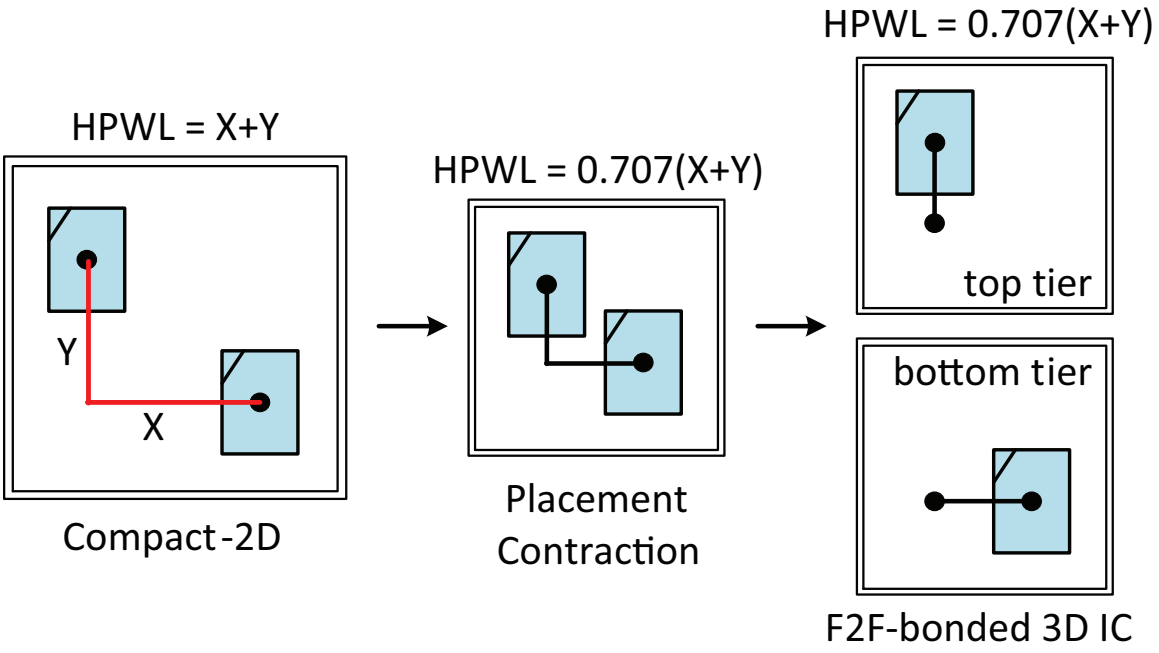


Figure 5.2: The need for interconnect RC scaling in a Compact-2D design. The length of interconnects will be reduced to 0.707X in the final F2F layout. In order to reflect this, we reduce the unit length RC to 0.707X in the Compact-2D design. The red line in the most left figure indicates an interconnect with reduced parasitics.

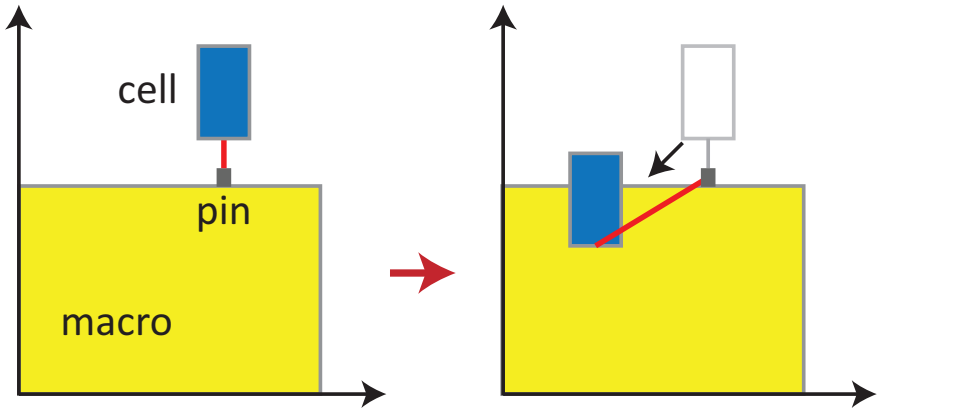
5.3.2 Placement Contraction

Once the Compact-2D design is implemented, the cell locations are linearly mapped to the 3D design footprint to finalize the optimal (X,Y) locations of cells in the F2F design. This is called placement contraction. Considering the linearly contracted HPWL of a net, the scaled interconnect RC parasitics of the Compact-2D design are the same as those of the F2F design in the original unit length RC. Also, it implies that the (X,Y) solutions based on the shrinking idea from S2D and interconnect RC scaling / placement contraction ideas from our C2D are ideally the same. However, C2D necessitates the P&R engines that handle the target technology node only while S2D relies on the CAD engines for the next technology node.

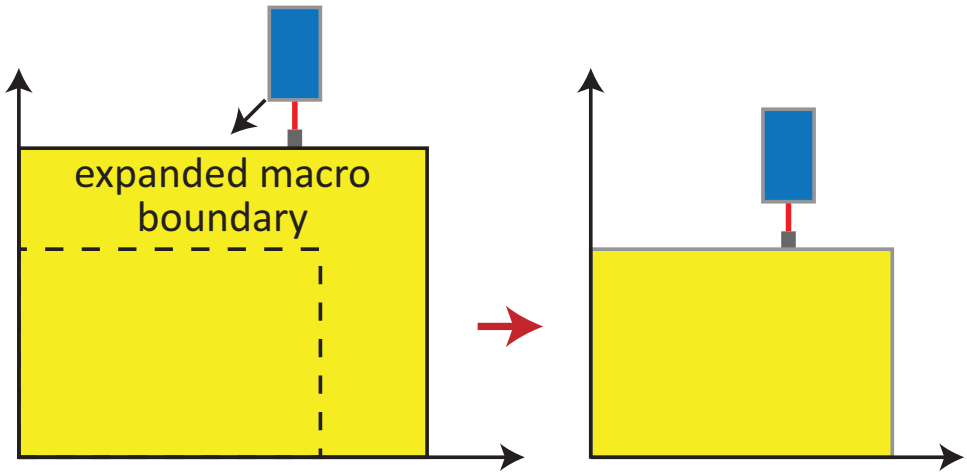
5.3.3 Handling Memory Macros

In the conventional 2D IC design, memory macros are preplaced in the floorplan without any overlaps, and none of standard cells is placed inside the memory macros. However, the Compact-2D design needs to allow overlaps of memory macros when they share the same (X,Y) location, but at different Z locations. Moreover, P&R engines should be allowed to place the standard cells inside the memory macro regions unless the memory macros fully occupy the regions in both tiers. Previously, S2D proposed shrinking the footprint of memory macros down to the minimum placement unit, and use placement blockages at its original boundary. Full placement blockages are used in the fully overlapped regions of memory macros, and restrict the standard cell placement. To enable the standard cell placement at partially vacated regions, 50% partial placement blockages are used. The pin locations of memory macros are retained, which serve as anchors for the standard cell placement regardless the footprint change of memory macros.

C2D follows the same way, but requires an additional step. Considering that the boundary of placement blockages should be the same as the original boundary of memory macros after placement contraction, the placement blockages for memory macros needs to be expanded by a factor of 1.414 for the Compact-2D design. The pin locations of memory macros also should be expanded to correctly anchor the standard cells around the placement blockages as shown in Figure 5.3. Therefore, at the floorplan stage of the Compact-2D design, we should prepare for the expanded memory macro Library Exchange Format (LEF) files (Memory Expansion), assign their tier locations, and preplace them manually considering the inter-module connectivity (Memory Preplacement), and generate placement blockages on the expanded memory regions while flattening the tier locations of memory macros (Memory Flattening).



(a) Contraction with the original macro pin location



(b) Contraction with the expanded macro pin location

Figure 5.3: The need for the expansion of memory boundaries in C2D flow. (a) The original macro pin location causes placement contraction to introduce unwanted routing change and cell overlap, (b) The macro boundary and its pin locations are expanded by a factor of 1.414 to resolve this issue.

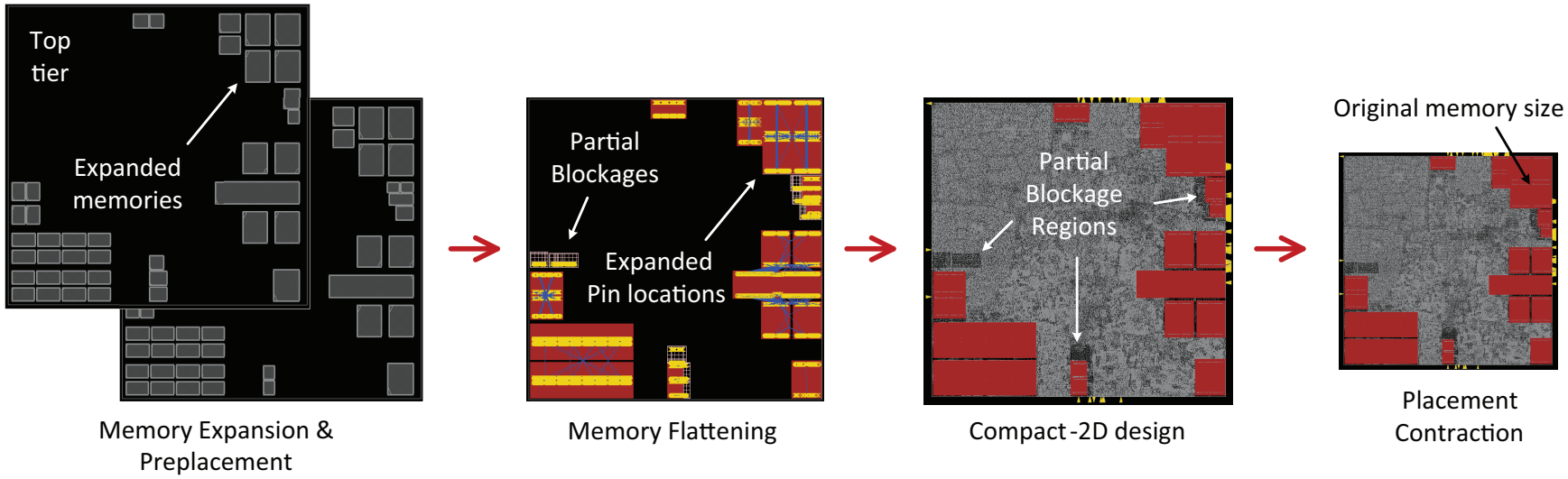


Figure 5.4: Our C2D flow demonstrated with OpenSparc T2 [56] single core design: memory expansion and preplacement, memory flattening, Compact-2D design, and placement contraction. Tier partitioning and Compact F2F via planning follow next.

5.3.4 Tier Partitioning

Since the tier locations of memory macros are preassigned manually, the standard cells within the memory macro boundaries move to the tier where memory macros do not occupy. To determine the Z location of each standard cell outside the memory macro boundaries, C2D introduces tier partitioning that utilizes bin-based placement-driven Fiduccia-Mettheyses (FM)-mincut partitioning algorithm [26]. Each partitioning bin is defined in a regular fashion on the final F2F footprint, and we run the algorithm based on the (X,Y) solution derived from placement contraction.

Bin-based placement-driven FM-mincut partitioning helps balance the area skew over the entire design footprint, otherwise resulting in huge white spaces or displacement from the optimal (X,Y) locations during placement legalization. The number of cutsize, which turns into the minimum number of inter-tier connections, is controlled by the size of partitioning bins. Too many cutsize leads to routing congestion, while too few cutsize decreases the power-performance benefits of F2F-bonded 3D ICs. Therefore, a sweet spot exists along the partitioning-bin size. Once tier partitioning determines the Z location of each cell, a placement engine legalizes the overlaps caused by placement contraction, and a Design Exchange Format (DEF) file for each die is created.

5.3.5 Compact F2F Via Planning

After we decide the (X,Y,Z) locations of standard cells, we should determine the F2F via locations. This is called F2F via planning. In this step, inter-tier 3D routing overhead, which is not accounted by the Compact-2D design, starts to affect the design closure. S2D is not only susceptible to this degradation, but none of 3D-routing-aware optimization is introduced after tier partitioning. In order to support post-tier-partitioning optimization (post-TP opt) to compensate the inter-tier routing overhead, C2D presents a unique stage named Compact F2F via planning. Compact F2F via planning consists of two steps, and following subsections describe them in detail.

Placement Row Splitting

Compact F2F via planning performs based on the 3D technology LEF which includes the definition and design rules of metal stacks in both tiers. 3D macro LEFs are required for the commercial router to distinguish the pin layer of macros based on their tier locations. Next, our in-house program creates a DEF and a Verilog file by instantiating the cells with 3D macro LEFs while flattening the tier location of cells. However, in order to fully utilize the optimization capabilities, the flattened DEF file should not have the placement overlaps although all synthesized gates are accommodated in the final F2F design footprint. Therefore, we split a placement row into the top and bottom rows, and change the height of standard cells in 3D macro LEFs to the half of the original to fit into the split rows.

In Figure 5.5, Row0 and Row1 are two adjacent placement rows to be split. Row1 is vertically flipped over to share the power rail with Row0. Now, placement row splitting turns each row into two horizontally split rows. In Row0, the bottom half is reserved for the bottom tier placement, and the top half for the top tier. However, in Row1, the bottom half is reserved for the top tier placement and the top half for the bottom tier due to the flipped orientation of Row1. As a result, the placement overlap is fully legalized while accommodating every cell in the design on the final F2F footprint. It is worth noting that the pin locations of standard cells are preserved regardless of splitting placement rows. Based on the retained pin locations and the same width of standard cells, accurate post-TP opt proceeds.

Post-Tier-Partitioning Optimization

C2D performs timing, power, and F2F via location co-optimization to close the design under inter-tier 3D routing overhead. Post-TP opt requires RC corners for the full 3D metal stack including the F2F via, and the timing corners for both top and bottom tiers. Thanks to placement row splitting, full optimization capabilities, including insertion, deletion, move, and resizing, are employed. Once the optimization is done, our in-house binaries create

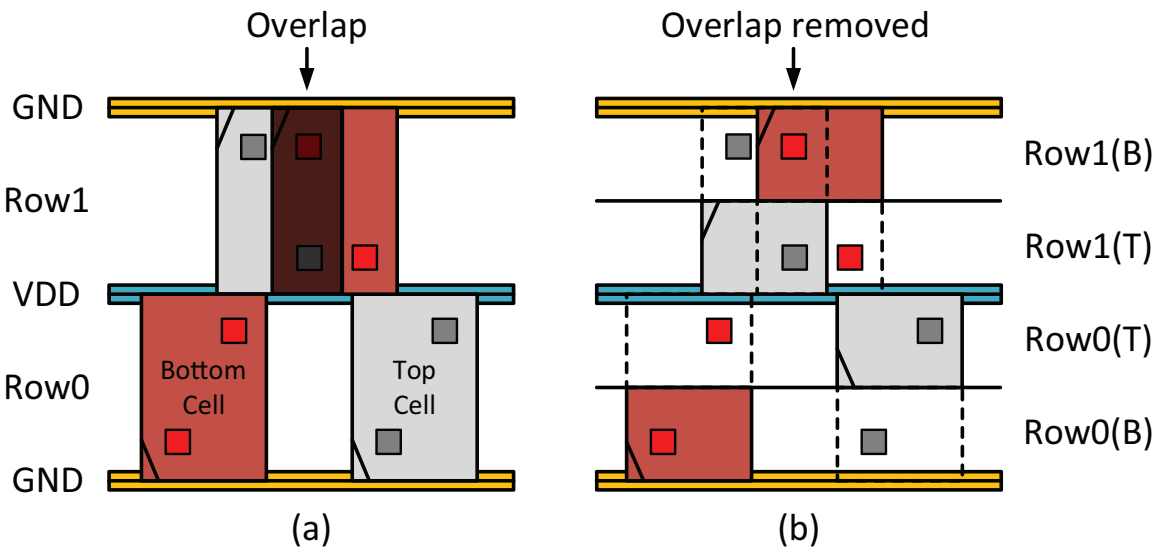


Figure 5.5: (a) Shrunk-2D flow [26] does not offer post-tier-partitioning optimization because of the placement overlap. (b) Placement row splitting in our C2D flow enables the optimization by fully legalizing the placement overlap.

a DEF file for each die that introduces the F2F vias as I/O ports (F2F ports) and contains the final cell locations by restoring the original cell height. Since the pin locations of cells are preserved regardless of having the cell height, we can easily retrieve the correct (X,Y) locations of cells based on the original cell height. Also, we generate a Verilog file for each die that presents the connectivity among F2F ports and the cells within a die. A top Verilog file that defines the connections between F2F ports in separate tiers is created, and lastly, we generate a top Standard Parasitic Exchange Format (SPEF) file that presents the RC parasitics of F2F vias.

5.3.6 Incremental Routing

Incremental routing is a CAD solution to preserve the routing result of Compact F2F via planning for the final GDSII file generation of each die. We first construct a graph for each net that consists of vertices and edges representing individual routing objects and their connectivities. Routing objects include a wire, a via, an I/O port, or a cell pin. The X/Y locations where those routing objects cross each other are kept along with their edge

definitions. Next, if the graph contains a F2F via vertex, we convert it into two vertices without an edge between them representing I/O ports for the top and bottom tier. Each vertex is only connected to the adjacent vertices that shares the same tier. As a result, the graph turns into a group of disconnected subgraphs, and each subgraph represents a 2D subnet on the specific die. Now, we reproduce the routing result for each subgraph based on the actual connection points defined in each edge. A depth-first search ordering is used to make the output in the format of DEF syntax. Finally, the routing information for each subgraph is delivered to the DEF for a corresponding die.

In the final GDSII file generation step, we use this routing information as an initial solution for sign-off physical design rule violation (DRV) fixing. The reason why DRV fixing is necessary is that tools built for 2D ICs do not support full DRV fixing for the pins outside the macro boundary while employing placement row splitting. When the sign-off DRV fixing is done, RC parasitics of each die are extracted, and we proceed the final 3D timing & power analysis.

5.4 State-of-the-art Comparison

In Table 5.2, we compare the timing & power savings of C2D with those of S2D based on the OpenSparc T2 [56] single core (SPC) design at 1.0GHz clock frequency. We use dual-Vt cell libraries in 28nm commercial-grade technology process design kit (28nm PDK). Six metal layers are used for 2D, and the top and bottom tiers for F2F implementations. The F2F via diameter, pitch, resistance and capacitance are assumed to be 0.5μm, 1.0μm, 0.5Ω, and 0.2fF, respectively. For the static power analysis, we set the switching activity as 0.1 for primary input ports and register output pins, and 2.0 for a clock port.

We observe that both C2D and S2D designs significantly decrease the net switching power thanks to the huge wirelength savings in F2F designs. Following buffer reduction contributes to the cell internal power savings. The total power reduction of C2D is 11.3% while S2D offers a 11.0% savings over 2D IC at iso-performance. In addition, it is remark-

Table 5.2: Timing & power comparison among 2D, S2D [26], and C2D using OpenSparc T2 [56] single core (28nm). Δ% shows % improvement over 2D. Target clock period is 1ns. C2D offers comparable power reduction and significant performance savings compared with S2D.

	2D	S2D	Δ%	C2D	Δ%
Total WL (m)	15.36	11.77	23.4%	11.55	24.8%
F2F Via #	-	154,127	-	193,487	-
Footprint (mm^2)	2.53	1.26	50.2%	1.26	50.2%
Total Power (mW)	338.20	300.87	**11.0%**	299.88	**11.3%**
Cell Power (mW)	82.12	79.11	3.7%	79.07	3.7%
Net Power (mW)	183.26	153.33	16.3%	150.86	17.7%
Leak. Power (mW)	72.83	68.43	6.0%	69.95	4.0%
Mem. Power (mW)	45.98	44.94	2.3%	44.77	2.6%
Comb. Power (mW)	171.30	140.90	17.7%	139.90	18.3%
Reg. Power (mW)	67.72	67.68	0.1%	69.80	-3.1%
Clk Tree Power (mW)	53.17	47.34	11.0%	45.40	14.6%
Worst Neg. Slack (ps)	-27.65	-52.52	**-89.9%**	-25.99	**6.0%**
Total Neg. Slack (ps)	-832.85	-846.94	**-1.7%**	-136.75	**83.6%**
Total Pos. Slack (ps)	35988.60	38884.50	8.0%	39422.20	9.5%

able that C2D reduces the total negative slack violations by 83.6% while S2D worsens the timing. This result not only shows that C2D offers comparable power reduction as the state-of-the-art S2D, but also proves that C2D builds timing-robust F2F designs. Most of all, C2D is more scalable than S2D in that our C2D flow performs with P&R engines, technology files, and design rules for the target technology and does not require handling of the next smaller node.

5.5 Experimental Results

In this section, we analyze the impact of each design step in C2D flow with LDPC, AES-128, and JPEG from OpenCore benchmark suites [41]. Assumptions on the technology and analysis are the same as Section 5.4 made. The initial utilization density for AES-128 and JPEG is 60%, while 40% for wire-dominated LDPC. The maximum clock frequency for each benchmark is 2.0GHz for LDPC, 5.4GHz for AES-128, and 2.16GHz for JPEG. Figure 5.6 shows the GDSII layouts of 28nm 2D and C2D-based F2F implementations for each benchmark including SPC at their maximum frequency.

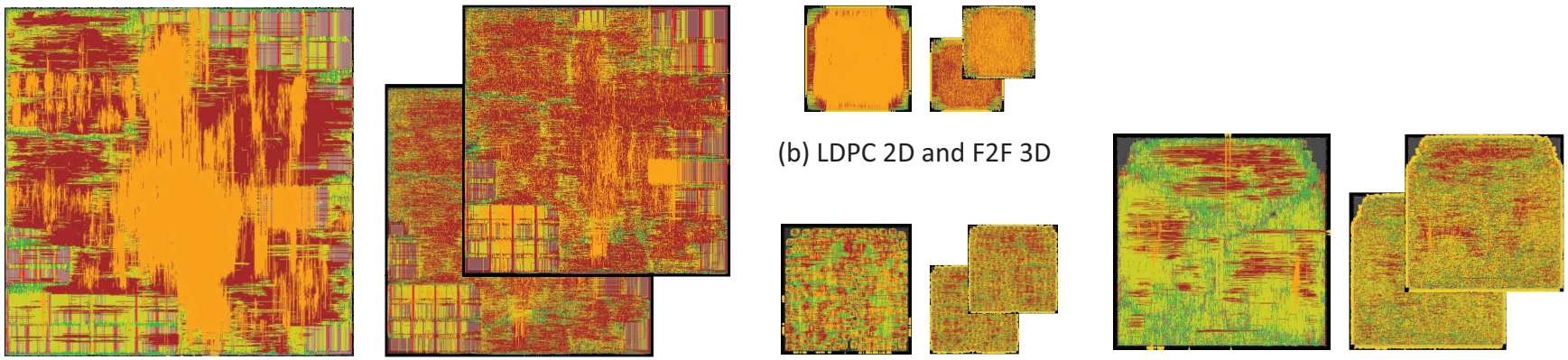


Figure 5.6: 28nm GDSII die images of 2D and F2F-bonded 3D implementations using our C2D flow. (a) SPC (1.0GHz), (b) LDPC (2.0GHz), (c) AES-128 (5.4GHz), (d) JPEG (2.16GHz).

5.5.1 Impact of Interconnect RC Scaling

In a Compact-2D design, we scale interconnect RC parasitics by a factor of 0.707 to imitate the parasitics of wirelength in the final F2F design based on that the footprint of the F2F design is exactly 50% of the 2D footprint. However, the RC scaling factor can be generalized and set to be 0.6 in case the F2F design footprint is 36% of the 2D footprint. Table 5.3 shows Compact-2D design results with various 3D/2D footprint ratios.

With a low RC scaling factor, such as 0.548, all benchmarks has huge power and standard cell area savings because of the reduced interconnect parasitics and the less number and lower drive-strength of buffers. However, since the target footprint is way smaller than the standard cell area savings, it results in the impractical placement utilization per each die in the F2F design. Assuming placement utilization in $[70\%, 80\%]$ range is allowed, our footprint savings reach up to 65% for LDPC, and 56% for both AES-128 and JPEG. In case of wire-dominated LDPC design, since the 2D footprint is determined by the routability, the huge wirelength reduction in the F2F design helps increase the footprint savings more.

When the same placement utilization in both 2D and F2F-bonded 3D ICs should be considered, we observe that 53-57% footprint savings are good target for all designs due to the buffer savings from the interconnect RC scaling. With a constraint on the exact 50% footprint savings, we find that 4-12% of placement utilization savings in F2F designs. In summary, sweeping the interconnect RC scaling helps to set the practical and optimal F2F design assumption. This also shows that C2D is incredibly flexible to design and find the optimal footprint of F2F designs for logic benchmarks thanks to the usage of original geometries for standard cells. For the rest of experiments, we keep the 50% footprint savings in F2F designs for all benchmarks to factorize the impact of other steps clearly.

5.5.2 Impact of Tier Partitioning

While placement contraction is deterministic in that the (X,Y) locations of cells are scaled by 0.707, bin-based tier partitioning is heuristic w.r.t the size of partitioning bins. Depend-

Table 5.3: Impact of target 3D footprint. Assuming placement utilization in $[70\%, 80\%]$ range is allowed, our footprint savings reach up to 65% for LDPC, and 56% for both AES-128 and JPEG

Footprint $(3D/2D)$	50%	45%	40%	35%	30%
RC Scaling	0.707	0.671	0.632	0.592	0.548
LDPC					
Std. Cell Area (mm^2)	0.180	0.178	0.177	0.172	0.169
3D Place. Util. per Die	58.31%	63.92%	72.03%	**79.69%**	91.29%
Place. Util $(3D/2D)$	87.83%	96.30%	108.50%	120.04%	137.51%
Total Power (mW)	179.23	174.48	167.70	158.03	153.85
Footprint $(3D/2D)$	50%	47%	44%	41%	38%
RC Scaling	0.707	0.686	0.663	0.640	0.616
AES-128					
Std. Cell Area (mm^2)	0.359	0.356	0.355	0.355	0.355
3D Place. Util. per Die	70.10%	73.88%	**78.99%**	84.58%	91.43%
Place. Util $(3D/2D)$	95.09%	100.22%	107.15%	116.15%	124.03%
Total Power (mW)	331.68	330.49	324.54	323.39	322.18
JPEG					
Std. Cell Area (mm^2)	0.943	0.941	0.939	0.936	0.933
3D Place. Util. per Die	70.71%	70.71%	**80.07%**	85.65%	92.16%
Place. Util $(3D/2D)$	96.03%	101.78%	108.73%	116.32%	125.15%
Total Power (mW)	579.17	573.52	565.84	563.80	560.10

ing on the partitioning-bin size, the number of cells applied to the algorithm varies, resulting in different cut sizes between the dies. Table 5.4 shows how the different partitioning-bin sizes change the number of 3D connections (F2F vias) and the wirelength of a design. F2F utilization indicates the F2F via usage out of the maximum available number of F2F vias inside the F2F design footprint. While the small bin size leads to the large number of F2F vias, the large bin size allows the algorithm to find the minimum cut size.

To explore the impact of the different number of 3D connections on the wirelength savings, a net is defined as either a 2D or a 3D net based on their F2F usage, and compare its wirelength with that in the Compact-2D design. We observe that the average wirelength per net is correlated to the optimal partitioning-bin size. If the bin size is way smaller $(5\mu m)$ than the average net wirelength, most of the nets become 3D, causing congestion and detour to meet the design rules for F2F vias. This is the reason that the wirelength savings of 3D nets decreases at $5\mu m$ bin, lowering the total wirelength savings. On the

other hand, if the bin size is too large, then most of the nets remain at 2D, requiring huge legalization caused by placement contraction. Therefore, embracing too much 2D nets again degrades the wirelength savings. LDPC shows the best wirelength savings (27.3%), which is almost ideal (29.3%), when the bin size is in the range of 20 to 80μm, while AES-128, and JPEG, which has short wirelength per net (gate-dominant), have 22.15% and 20.47%, respectively at 10μm bin. It is noteworthy that gate-dominant circuits steeply loose the wirelength savings along with increasing the bin size over the sweet spots. We determine the size of partitioning bins as 40μm for LDPC, 10μm for both AES-128 and JPEG, and proceed Compact F2F via planning.

5.5.3 Impact of Compact F2F Via Planning

Using LDPC, Table 5.5 demonstrates that how negatively inter-tier 3D routing affects the timing, and how effectively post-TP opt fixes the timing violations. Since the Compact-2D design does not account the inter-tier routing overheads when it is implemented, we observe that the worst negative slack (WNS) is degraded to 5.87x, and 7.71x for the total negative slack (TNS) after the inter-tier 3D routing is done. All of these timing violations are fixed after we perform post-TP opt. The WNS is improved by 44.4% and the TNS is restored by 91.5% with the negligible power overhead. This proves that post-TP opt in Compact F2F via planning is critical to implement timing-robust F2F designs. In general, post-TP opt restores the timing by inserting or up-sizing the buffers while minimizing the power increase. However, if the power overhead becomes the issue, then post-TP opt can start to delete or down-size the buffers at the expense of the timing margin.

5.5.4 Impact of Incremental Routing

Table 5.6 shows how final DRV fixing and tier-by-tier 2D routing affects the design result from post-TP opt, and how much better our incremental routing performs than the existing iterative tier-by-tier routing method in S2D. Iterative routing starts the tier-by-tier routing

Table 5.4: Impact of tier partitioning bin size. Smaller bins cause more F2F vias to be used and tend to improve WL saving for 3D. Saving values are w.r.t. 2D results.

Bin Size (μm)	5	10	20	40	80
LDPC					
Bin #	6,169	1,542	386	96	24
Avg. Cell # / Bin	11	42	169	677	2,707
F2F Via #	**55,468**	**26,999**	**20,850**	**19,802**	**19,726**
F2F Util. (%)	34.20	16.65	12.86	12.21	12.16
Avg. WL / net (μm)	39.16	38.85	38.83	38.84	38.82
3D Net # (%)	61.41	24.71	17.73	16.89	16.75
3D Net WL Savings (%)	26.73	27.58	27.87	27.87	27.95
2D Net WL Savings (%)	26.60	26.93	26.80	26.79	26.77
Total WL Savings (%)	**26.70**	**27.28**	**27.32**	**27.30**	**27.33**
AES-128					
Bin #	10,247	2,562	640	160	40
Avg. Cell # / Bin	14	55	219	877	3,507
F2F Via #	**104,306**	**61,902**	**51,460**	**22,311**	**10,824**
F2F Util. (%)	39.16	23.24	19.32	8.38	4.06
Avg. WL / net (μm)	16.45	16.24	16.56	18.16	18.83
3D Net # (%)	59.67	28.11	22.91	11.14	5.96
3D Net WL Savings (%)	20.57	22.10	21.50	18.45	16.73
2D Net WL Savings (%)	22.74	22.20	19.95	11.46	8.76
Total WL Savings (%)	**21.14**	**22.15**	**20.60**	**12.94**	**9.71**
JPEG					
Bin #	26,680	6,670	1,668	417	104
Avg. Cell # / Bin	11	43	171	682	2,729
F2F Via #	**240,301**	**120,921**	**94,868**	**71,353**	**53,810**
F2F Util. (%)	35.17	17.70	13.88	10.44	7.88
Avg. WL / net (μm)	14.54	14.57	14.76	15.06	15.67
3D Net # (%)	61.36	25.19	18.42	13.27	10.10
3D Net WL Savings (%)	20.69	21.73	21.61	21.39	19.11
2D Net WL Savings (%)	19.76	19.01	17.66	15.53	12.17
Total WL Savings (%)	**20.47**	**20.31**	**19.29**	**17.60**	**14.28**

from scratch on top of the placement result of post-TP opt. This leads to a different routing result from post-PT opt due to the final DRV fixing, and perturbs the design closure. We observe that the worst negative slack is degraded to 1.86x, and 25.88x for the total negative slack after using iterative routing. However, our incremental routing preserves the worst negative slack in the acceptable level (less than 25ps under 0.5ns clock period), and retains the total wirelength and power results close to the optimization result (less than 1% overheads).

Table 5.5: Impact of post-tier-partitioning optimization. Δ% indicates its savings. Inter-tier 3D routing (A vs. B) introduces huge timing violations, and our optimization (B vs. C) fixes the timing violations with the negligible power overhead.

Design	LDPC			
Stage	Before 3D Routing (A)	After 3D Routing NO-Opt (B)	After 3D Routing YES-Opt (C)	Δ%
Total Cell (#)	65,187	65,187	65,271	-0.1
Worst Neg. Slack (*ps*)	-7.42	-43.57	-24.23	**44.4**
Total Neg. Slack (*ps*)	-341.86	-2637.13	-222.99	**91.5**
Total Pos. Slack (*ps*)	19194.40	17042.80	27072.40	**58.8**
Violated Path (#)	20	383	27	**93.0**
Total Power (*mW*)	179.23	178.25	178.49	**-0.1**

Table 5.6: The impact of Final DRV fixing and tier-by-tier 2D routing after post-TP opt. We note that the incremental routing (Incr-R) used in C2D preserves the timing closed by post-TP opt (A vs. C) better than the iterative routing (Iter-R) in S2D [26] (A vs. B). Incr-R also offers smaller wirelength and power overheads for the tier-by-tier routing than Iter-R. Δ% indicates the savings from Incr-R over Iter-R.

Design	LDPC			
Stage	Before 2D Routing (A)	After 2D Routing Iter-R (B)	After 2D Routing Incr-R (C)	Δ%
Total WL (*m*)	2.721	2.754	2.750	**0.1**
Worst Neg. Slack (*ps*)	-24.23	-45.17	-25.16	**44.3**
Total Neg. Slack (*ps*)	-222.99	-5771.74	-1599.73	**72.3**
Total Pos. Slack (*ps*)	27072.40	11257.00	15107.10	**34.2**
Violated Path (#)	27	734	402	**45.2**
Total Power (*mW*)	178.49	179.53	179.15	**0.2**

5.5.5 Runtime Analysis

In Table 5.7, we tabulate runtime for each design step of 2D, S2D, C2D flows to build LDPC, AES-128, and JPEG. Intel(R) Xeon(R) CPU E5-2640 @ 2.50GHz is used, and 16 cores are employed while running Cadence Innovus. Thanks to the reduced interconnect loads and HPWL savings, Compact-2D designs take 34% less time than 2D until the post-route optimization is done (Compact-2D designs take 10% less time than Shrunk-2D designs at best). However, the total runtime of C2D is longer than that of 2D by a maximum 50% (JPEG), due to the additional steps starting from placement contraction. Although incremental routing achieves a huge runtime savings up to 60% compared with iterative

routing in S2D, post-TP opt takes a large portion of the F2F design flow, resulting in 21% runtime overhead in C2D over S2D at worst.

Table 5.7: Runtime comparison (in minutes): Intel(R) Xeon(R) CPU E5-2640 @ 2.50GHz, 16 cores usage for Cadence Innovus run.

Design	LDPC			AES-128			JPEG		
Runtime (*min*)	2D	S2D	C2D	2D	S2D	C2D	2D	S2D	C2D
Placement	3	3	3	3	3	3	7	7	7
Pre-CTS Opt.	44	19	22	33	29	28	59	54	55
Clock Tree Syn.	3	5	3	5	6	5	15	17	13
Post-CTS Opt.	8	6	6	12	9	7	15	12	12
Routing	6	8	6	5	7	5	9	11	8
Post-route Opt.	11	10	10	8	8	8	20	19	19
Place. Contr.	-	-	1	-	-	1	-	-	2
Tier Part.	-	1	1	-	3	3	-	11	11
F2F Via Plan.	-	10	10	-	10	10	-	19	19
Post-TP Opt.	-	-	20	-	-	15	-	-	39
Iter. Routing	-	11	-	-	12	-	-	20	-
Incr. Routing	-	-	7	-	-	7	-	-	11
Signoff Analysis	2			3			10		
Final Total	**77**	**75**	**91**	**69**	**90**	**95**	**135**	**180**	**206**

5.5.6 Commercial 2D vs. C2D

Based on the optimal footprint derived from Section 5.5.1, we compare the design results of commercial 2D with C2D-based F2F designs. The total area savings of F2F designs over the 2D is 57.8% for LDPC, and 53.0% for both AES-128, and JPEG. As shown in Table 5.8, our C2D flow offers a 20-34% wirelength savings and a 4-13% standard cell area savings. Therefore, the wire-dominated LDPC, which shows the highest wirelength to standard cell area ratio, benefits most from C2D in terms of the total power savings at iso-performance (26.8%), whereas the lowest wirelength to standard cell area ratio benchmark, JPEG, gains the lowest total power savings (5.7%). An interesting trend is that the standard cell area reduction depends on the ratio of sequential cell count to the total number of cells. Since the number of sequential cells in a design is not changed, only reduced drive strength for the sequential cells contributes to the power savings. On the other hand, buffers are

optimized in both number and strength. Therefore, LDPC, which has the lowest sequential cell count to the total cell count ratio (2.7%), achieves the largest standard cell area savings (12.7%) in the F2F design.

Table 5.8: Iso-performance power comparison between commercial 2D vs. C2D. Δ% indicates the savings over 2D designs.

Design	2D	C2D	Δ%
LDPC, 2GHz			
Footprint ($\mu m \times \mu m$)	555.7 × 555.1	361.2 × 360.8	57.8
F2F Via Count	-	21,575	-
Cell Count	77,024	64,610	16.1
Seq. Cell Count (%)	2,048 (2.7%)	2,048 (3.2%)	0.0
Standard Cell Area (μm^2)	204,782	178,876	**12.7**
Total Wirelength (m)	3.8	2.5	**33.6**
Tot. WL / Cell Area (m^{-1})	**18.7**	14.2	24.1
Switching Power (mW)	193.9	136.9	29.4
Cell Internal Power (mW)	33.0	28.8	12.7
Leakage Power (mW)	11.1	8.2	26.1
Total Power (mW)	237.8	174.0	**26.8**
AES-128, 5.4GHz			
Footprint ($\mu m \times \mu m$)	716 × 715.6	490.9 × 490.6	53.0
F2F Via Count	-	63,211	-
Cell Count	147,483	140,960	4.4
Seq. Cell Count (%)	10,688 (7.2%)	10,688 (7.6%)	0.0
Standard Cell Area (μm^2)	377,702	361,096	**4.4**
Total Wirelength (m)	2.9	2.2	**22.9**
Tot. WL / Cell Area (m^{-1})	**7.7**	6.2	19.5
Switching Power (mW)	250.8	223.7	10.8
Cell Internal Power (mW)	113.6	108.4	4.6
Leakage Power (mW)	17.5	16.1	8.0
Total Power (mW)	381.9	348.2	**8.8**
JPEG, 2.16GHz			
Footprint ($\mu m \times \mu m$)	1156.3 × 1153.7	792.8 × 791.0	53.0
F2F Via Count	-	121,357	-
Cell count	312,451	284,884	8.8
Seq. Cell Count (%)	37,538 (12.0%)	37,538 (13.2%)	0.0
Standard Cell Area (μm^2)	982,231	943,812	**3.9**
Total Wirelength (m)	5.8	4.6	**20.2**
Tot. WL / Cell Area (m^{-1})	**5.9**	4.9	16.9
Switching Power (mW)	415.8	385.9	7.2
Cell Internal Power (mW)	195.1	189.9	2.7
Leakage Power (mW)	30.2	28.5	5.6
Total Power (mW)	641.1	604.4	**5.7**

5.6 Conclusions

To maximize the utilization of 3D interconnect and the power-performance-area benefit of F2F-bonded 3D ICs, in this research, we proposed a full-chip RTL-to-GDSII physical design solution named Compact-2D (C2D) that offers a commercial-quality F2F-bonded 3D IC physical layout. We presented interconnect RC scaling, placement contraction, and memory expansion idea, which allows us to utilize the original technology files and design rules of the target technology node for a F2F-bonded 3D IC implementation. We also introduced placement row splitting idea to enable post-tier-partitioning optimization in our C2D flow, which is completely missing in the state-of-the-art F2F physical design solution. With our extensive experiments and analysis, we evaluated the impact of those ideas in the final F2F design results, and showed that using 28nm process design kit, F2F-bonded 3D ICs implemented by our C2D flow offers a maximum 26.8% of total power reduction with a maximum 15.6% silicon area savings compared to the 2D IC designs at iso-performance.

CHAPTER 6

DESIGN AND ARCHITECTURAL CO-OPTIMIZATION OF MONOLITHIC 3D LIQUID STATE MACHINE-BASED NEUROMORPHIC PROCESSOR

The liquid state machine (LSM) [57] is one model of recurrent spiking neural networks (SNNs). It is constructed with a recurrent reservoir that consists of a set of randomly connected spiking neurons, and an output layer which receives reservoir responses as inputs. The reservoir synapses are fixed in the standard LSM model to relax the difficulty in training. The reservoir exhibits complex non-linear dynamics and acts as a pre-processor mapping input patterns to a higher-dimensional transient response, which is fed to the output neurons for final classification through the trainable synapses, referred to as output synapses. The LSM is especially competent for spatio-temporal pattern classification such as speech recognition.

While SNNs hold a lot of promise due to their bio-plausibility and hardware implementation efficiency, the training of SNNs still remains challenging. It is difficult to develop a powerful gradient-based learning mechanism for SNNs, particularly recurrent SNNs. To this end, the LSM is envisioned as a good tradeoff between the ability in tapping the computational power of recurrent SNNs and engineering tractability. Recently, cost-effective hardware implementations of the LSM have been investigated, along with bio-inspired training algorithms to tune both the reservoir and output layer. For example, [25] proposed a supervised probabilistic spike-dependent output tuning algorithm, [58] proposed an LSM-based learning processor with runtime programmable arithmetic precision and data-dependent reconfiguration, and [59] proposed a self-organizing LSM architecture with hardware-friendly spike-timing-dependent-plasticity rules for reservoir tuning.

The synergetic impact of M3D integration with LSM architecture is worth to note. M3D offers great benefits in neural network designs due to the neuromorphic architecture with a

huge number of connections at both intra-neuron and inter-neuron levels. In this work, for the first time, we explore the design methodologies and study the benefits offered by M3D ICs in LSM-based speech recognition processors.

The major contributions of this study are (1) We carry out ASIC design for LSM neural processors in 2D and M3D IC with detailed design comparison. (2) We explore the impact of different synapse models and memory distributions on the power-performance-area-accuracy benefit of M3D LSM neural processors. (3) We conduct vector-based functional verification and power-performance-area-accuracy analysis for the real-world task of speech recognition.

6.1 LSM Architecture Description

6.1.1 Processor Architecture

The reservoir and output layer are realized by a reservoir unit (RU) and a training unit (TU), respectively, and each neuron is implemented by a digital processing unit. The overall LSM processor architecture is adopted from [59], and there are 135 digital reservoir neurons (RNs) and 26 digital output neurons (ONs) as depicted in Fig. 6.1. External input spikes are fed to their targeted reservoir neurons through the crossbar interface with a predefined connectivity pattern. The spikes generated from reservoir neurons are registered (i.e. Reservoir spike buffer[134:0]) and propagate to the TU. Meanwhile, these spikes are also sent back to other reservoir neurons in the RU through reservoir crossbar interface. The operations of neurons at the same layer are executed in parallel under the control of a global finite state machine (FSM).

The on-chip training of the LSM processor can be divided into two phases. First, during the reservoir training phase, the RU is trained by a hardware-friendly spiking-timing dependent plasticity (STDP) algorithm [59] until its synaptic weight distribution converges. Then, a bio-plausible supervised spike-based learning algorithm [60] is employed on the TU for the main classification function. In this second phase, the reservoir maintains its

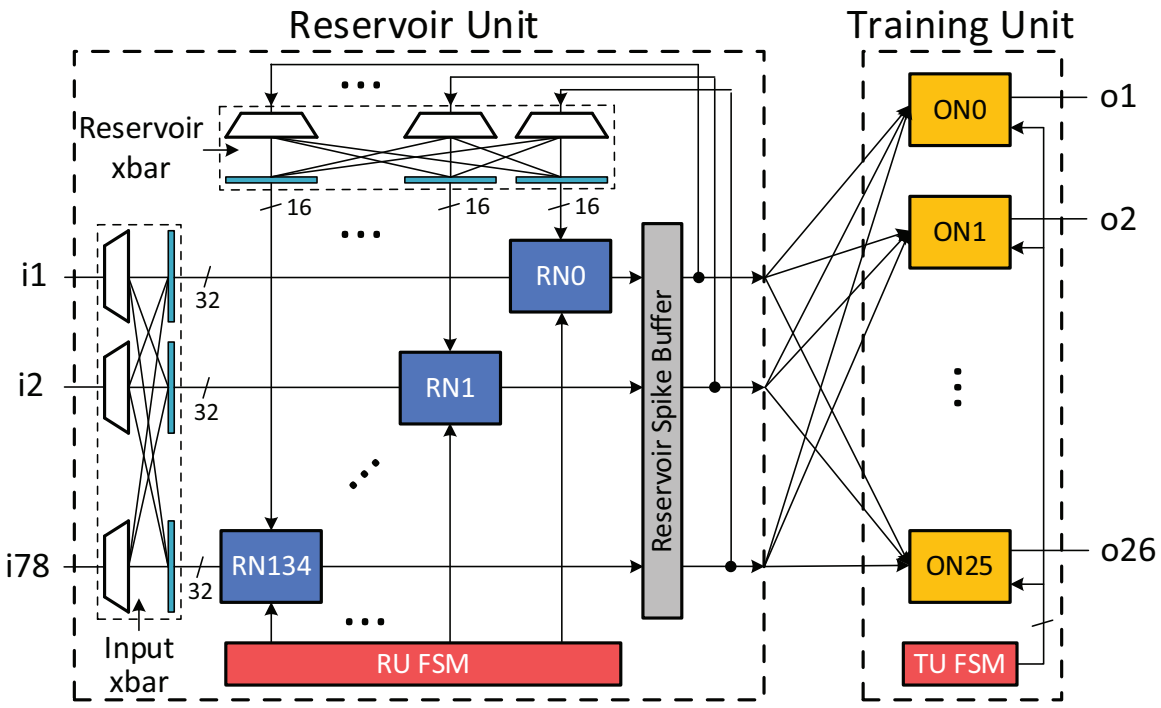


Figure 6.1: Our LSM-based neuromorphic processor architecture. There are 135 reservoir neurons (RNs) in the reservoir unit, and 26 output neurons (ONs) in the training unit. Each RN receives up to 32 external input spikes and up to 16 pre-synaptic reservoir spikes. Each ON has a full connection to the individual RNs to receive the reservoir response.

synaptic weights while producing spike responses to the TU.

6.1.2 Digital Spiking Neuron Implementation

The proposed LSM neural processor operates through a series of computational steps that are controlled by the corresponding states of the global finite state machine (FSM) in the RU and TU respectively, and involve a number of logic cells and storing elements inside each neuron. Based on the architectural and functional properties, we partition the implementation of a single digital neuron into three functionally dependent modules: the synaptic input processing module, the spike generation module, and the learning module. At each emulation time step, these three modules activates in order, controlled by the well-defined states of the global finite state machines at the reservoir and output layer.

The synaptic input processing module computes synaptic responses upon arrival of

spike inputs. As a baseline, we implement second-order dynamic synaptic model [25], in which the excitatory and inhibitory synapses have their separate state variables:

$$\begin{aligned} EP(t+1) &= EP(t)(1-1/\tau_{EP}) + \sum w_i \cdot S_+(i) \\ EN(t+1) &= EN(t)(1-1/\tau_{EN}) + \sum w_i \cdot S_+(i) \\ IP(t+1) &= IP(t)(1-1/\tau_{IP}) + \sum w_i \cdot S_-(i) \\ IN(t+1) &= IN(t)(1-1/\tau_{IN}) + \sum w_i \cdot S_-(i) \end{aligned} \tag{6.1}$$

where $EP(t+1)$ ($EP(t)$) and $EN(t+1)$ ($EN(t)$) are excitatory state variables of a neuron at the $(t+1)$th (tth) biological time step, while IP and IN are inhibitory ones. τ_{EP}, τ_{EN}, τ_{IP}, τ_{IN} are the decay time constants of the corresponding state variables, w_i is the synaptic weight and S_i is the spike of the i-th synapse.

When updating the state variables in a neuron, the input synapses are examined in serial. If there is an input spike at the current time step, the synaptic weight of the associated synapse will be added to the corresponding state variables. After the synaptic responses are generated, the spike generation module updates the membrane potential V_{mem} with the response based on the widely used leaky integrate-and-fire (LIF) model and generates a spike if the membrane potential exceeds a pre-defined threshold. The calculation of neuron membrane potential update follows below.

$$V_{mem}(t+1) = V_{mem}(t)(1-1/\tau_m) + \frac{EP-EN}{\tau_{EP}-\tau_{EN}} - \frac{IP-EN}{\tau_{IP}-\tau_{IN}} \tag{6.2}$$

where $V_{mem}(t+1)$ ($V_{mem}(t)$) is the membrane potential at the $(t+1)$th (tth) biological time step, τ_m is the decay time constant of membrane voltage.

At last, the learning module activates in each emulation time step after the spike generation module finishes the process and tunes the afferent pre-synaptic weights of the associated neurons with a bio-inspired supervised spike-based algorithm [25]. In our LSM neural processor, we implement the activity-dependent clock gating adopted from [60] and

directly gate on the clock signals inside each neuron. The clock signal of each functional module only toggles when the module needs to be activated.

6.2 Design Flow and Methodologies

6.2.1 Baseline RTL-to-GDS Flow

In this work, we implement full-chip RTL-to-GDSII ASIC LSM neural processors using commercial 28nm process design kit at the block-level with 135 reservoir neurons and 26 output neurons to reduce the design complexity and facilitate IP reuse. While using the conventional hierarchical design flow for 2D IC design, we adopt Shrunk-2D flow [26], and extend it to build the top-down hierarchical M3D IC design. The diameter of an MIV used is 50nm and the RC parasitics are (10Ω, 0.2fF) based on 28nm PDK metal pitches, via-sizes, and via aspect ratio.

6.2.2 Hierarchical Shrunk-2D

We carry out two-level folding where each individual neuron is partitioned into two tiers, and top-level cells are partitioned into two tiers incrementally. For our hierarchical LSM neuromorphic processor design, we first decide the top-level floorplan based on the shrunk layout geometry, and derive the timing budget for the reservoir and output neuron blocks. Based on these block timing constraints, we follow the Shrunk-2D flow for each neuron, and build two-tier folded M3D neuron designs. To build top-level Shrunk-2D design, we use Shrunk-2D design for individual neuron blocks. Although the top-level Shrunk-2D design finds the neurons unfolded, the individual neuron is actually folded, and fully occupy the placement area in both tiers. Therefore, we need to split the Shrunk-2D neuron blocks into two different blocks that share the same X,Y location but placed on the separate tiers. This is called neuron splitting.

The top-level netlist and placement result also should be updated in accordance with the neuron splitting. Then, we build the top-level M3D design. To avoid routing perturbation

derived from the MIVs inside the M3D neurons during top-level MIV planning, we use abstract macro LEFs that do not contain the MIV ports. Once the tier-by-tier routing for the top-level is done, we replace the neuron macro LEFs into the ones with MIV ports, and revise the Verilog and routing results to support full connectivity including both top-level and neuron-level 3D connections. This is called neuron MIV port punching. Lastly, we generate GDSII file for our M3D LSM neuromorphic processer, and proceed the signoff M3D timing and power analysis.

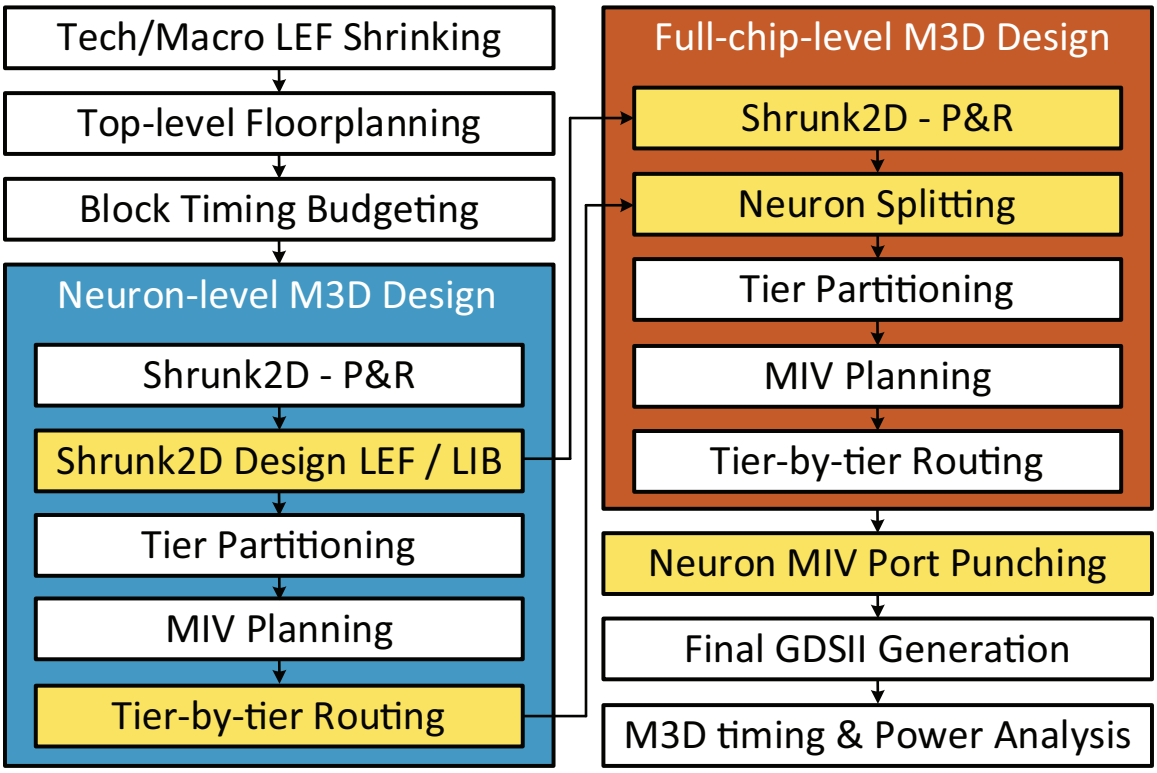


Figure 6.2: Our hierarchical Shrunk-2D flow to enable two-level design folding: individual neuron is partitioned into two tiers, and top-level design is also tier partitioned.

6.2.3 Design Methodology Enhancements

We use six metal layers in 2D IC while only four metal layers are allowed inside each neuron. For M3D IC, four+four metal layers are used inside the folded neuron to provide the same routing resources as the 2D neuron, and additional two routing layers on the

top tier are dedicated to the inter-neuron routing. In a reservoir neuron, we use flip-flops to store synaptic weights considering relatively limited pre-synaptic fanins. For output neurons, however, we use register-file modules to store the weights since they have trainable synapses in full connection to the reservoir unit. Memory modules are generated using a commercial memory compiler for the used 28nm technology node, and occupy up to four metal.

We fold each individual neuron block into two tiers and partition the cells and pins of the neuron block to maximize the area and power benefit leveraged from M3D IC. For reservoir neurons, we first put all functional cells in synaptic input processing module and action potential (spike) generation module on the top tier so that they are on the same layer with the global nets and closer to the external connections to package pins. Then, we separate the 16-bit reservoir spike input pins into two groups and put the 8 lower bits of the reservoir spike inputs and their peripheral logic cells on the bottom tier. All other input and output pins are assigned to the top tier for simplicity. Since the reservoir spike input pins are connected to the synaptic input processing module, by having half of the reservoir spike inputs on the bottom tier, we increase the vertical connections inside each neuron.

The memory inside each output neuron takes a large part of the layout. Considering that the routing across the memory is costly, we put the memory and its peripheral logic cells on the bottom tier while all other cells (i.e. synaptic input processing module, action potential (spike) generation module and the learning module) on the top tier. Similar to the reservoir neuron, we also partition the spike input pins of an output neuron into two evenly sized groups and put one group on the bottom tier to increase the vertical connections.

M3D neurons are arranged in a floorplan similar to the 2D IC layout, but with each neuron smaller in footprint and spread across both the tiers as demonstrated in Figure 6.3. The two-tier M3D IC floorplan footprint is half that of 2D IC. Therefore, the total silicon area used is the same. Since output neurons communicate with all reservoir neurons, the 26 output neurons are uniformly arranged in the center of the floorplans.

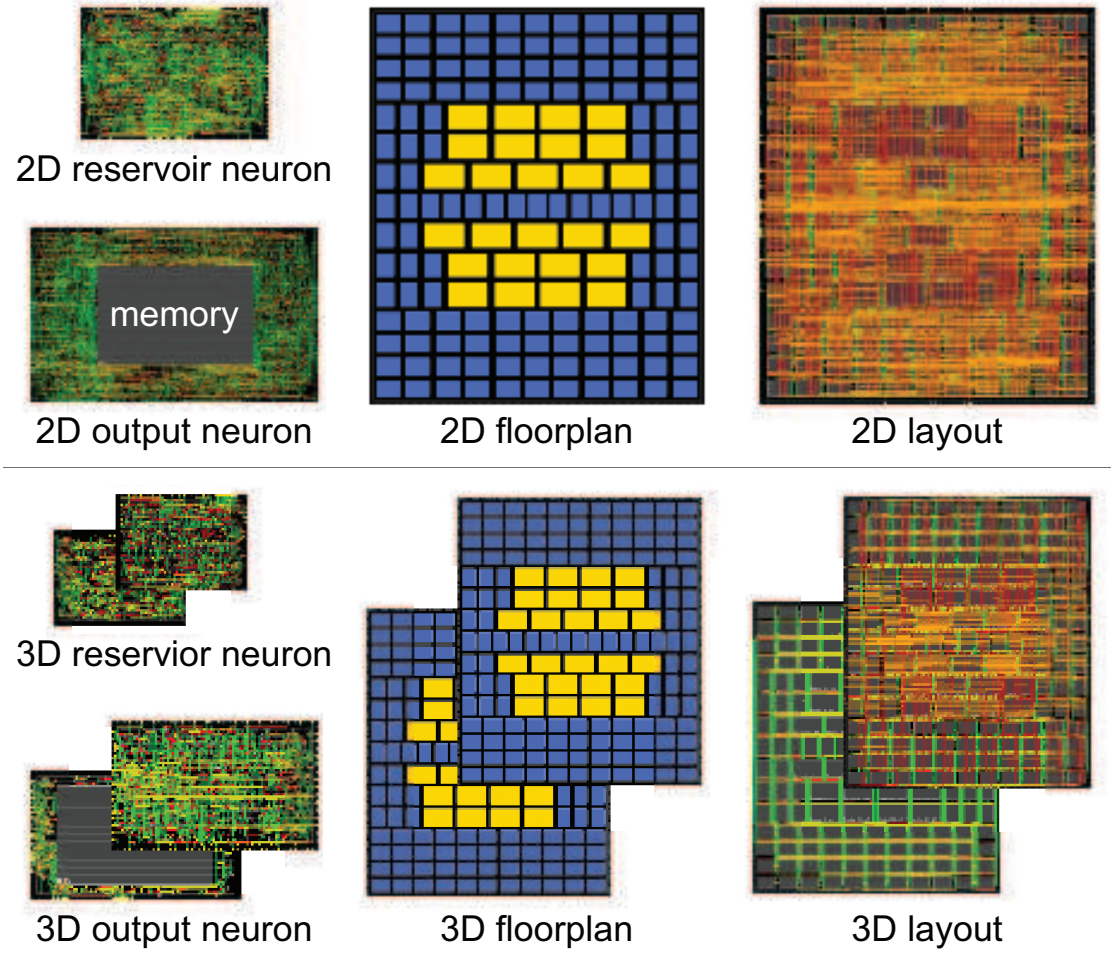


Figure 6.3: 2D vs. M3D designs of reservoir neuron, output neuron, and full-chip. Reservoir neurons are in blue, and output neurons in yellow in the flooplan.

6.3 Design/Architecture Co-Design

6.3.1 Memory Sharing

In the proposed LSM processor, a large number of memory resources are required for weight storage, thus an efficient memory design scheme is important for the hardware cost and energy efficiency. The straightforward way is to distribute the memory module inside each neuron. The depth of the memory depends on the number of pre-synapses of the neuron, which is set to be 16 for reservoir neurons and 135 for output neurons. The

memory width represents the synaptic weight bit resolution, which is 2 and 8 respectively for reservoir and output synapses.

Although the distributed memory architecture is easy to implement, it results in large peripheral overhead due to a large number of memory modules. To improve the memory efficiency, we replace the individual weight storage inside the neuron with a large shared memory at reservoir and output layer, respectively. This is based on that, at each emulation time step, all neurons at the same layer work in parallel; The synaptic weights are accessed in serial following the same order based on their index. This indicates that, in any state, the neurons at the same layer are actually accessing the same address of their own memory, though the values stored at that address might be different. Given that, in the shared memory architecture, we store all synaptic weight values in a row that are previously at that same address in the distributed memory, and the values are associated with different neurons by the bit index. When updating the weight value, the updated synaptic weights from all neurons will first be concatenated to one word then write to the intended address. When reading the weights, different parts of the memory output are assigned to their targeted neurons.

6.3.2 Synaptic Model Complexity Reduction

Reducing synaptic model from the second-order dynamics to the first-order dynamics is another approach to optimize the overall power-performance-area-accuracy benefit. In the first-order synapse model, there is only one state variable E in each neuron, which represents the overall synaptic response among all its input spikes:

$$E(t+1) = E(t)(1 - 1/\tau_E) + \sum_i w_i \cdot S_i \tag{6.3}$$

where $E(t+1)$ ($E(t)$) is the first-order state variable at the $(t+1)$th (tth) biological time step, τ_E is the decay time constant of the synaptic response.

The calculation of neuron membrane voltage update in the first-order synaptic model is also different from the neuron that is in the second-order:

$$V_{mem}(t+1) = V_{mem}(t)(1 - 1/\tau_m) + \frac{E}{\tau_E} \tag{6.4}$$

In the following sections, we will show that these two approaches will effectively reduce the area and power of the M3D LSM neural processor without hurting the classification accuracy too much.

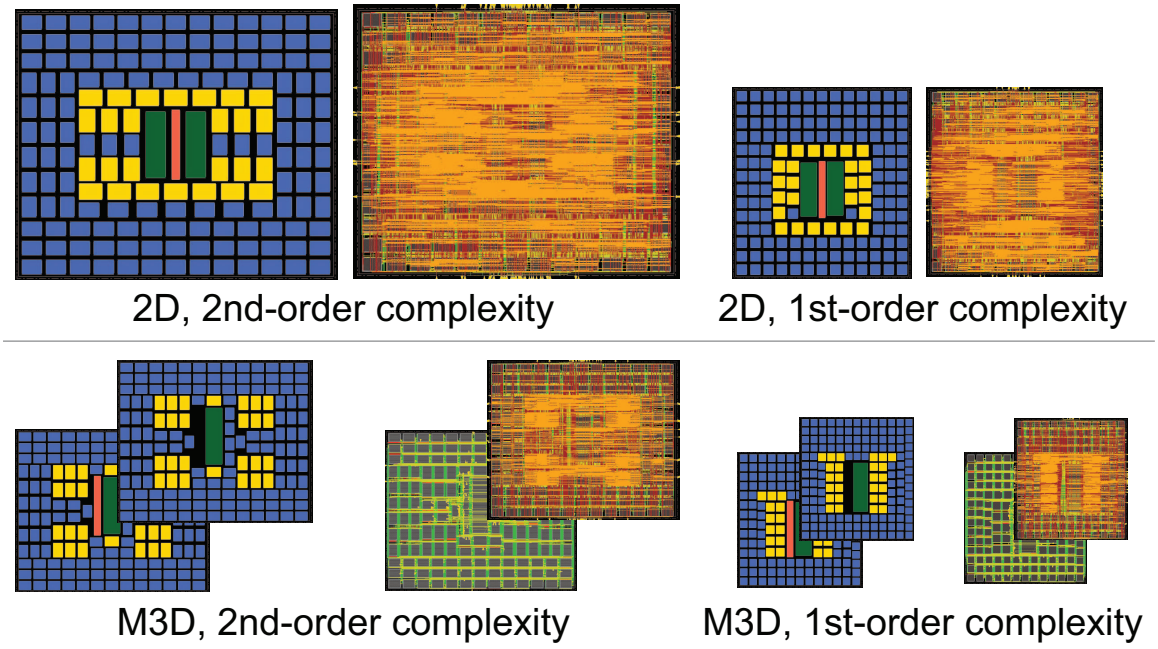


Figure 6.4: 2D vs. M3D LSM processors with memory sharing & synaptic model complexity reduction schemes. In red is shared memory for the reservoir neurons (yellow), and in greens are for output neurons (blue).

6.3.3 Individual Neuron Results

First, we compare the 2D neuron designs of the shared memory architecture to those of the baseline distributed memory 2nd-order synapse model architecture. The distributed memory modules occupy huge placement area and internal power inside the individual neuron. Using shared memory architecture, these modules are now located at the top-

level hierarchy, and it leads to 14% and 54% footprint area savings for reservoir and output neurons, respectively. The reduced number of flip-flops and the absence of memory module allows to 24%, and 48% internal power savings, and reduced footprint leads to 15%, and 23% switching power savings in the reservoir and output neuron, respectively. For output neuron, eliminating the memory module not only helps to reduce the huge internal power, but also removes the routing blockage over the memory module, resulting in the efficient routing.

On top of this huge benefit, reducing synapse model complexity enables more compact neuron design by reducing the cell count from the relaxed synaptic weight precision. This results in 57% and 75% footprint savings from shared memory first-order synapse model architecture compared to the baseline architecture, and 65% and 69% of total power savings for the reservoir and output neuron, respectively.

We observe that M3D designs offer even more savings in terms of footprint, and power consumption for all neuron designs on top of the architectural optimization benefit. Assuming no silicon area overhead, 50% additional footprint savings of M3D design lead to additional 9% and 4% total power savings in the reservoir neuron and 15% and 4% for output neurons in two-different architectures. It is note worthy that the shared memory second-order syanpse model architecture maximizes the M3D power benefits in both neuron designs. This is because, targeting 1GHz, the neurons of first-order synapse model architecture have large timing margin in the path, and meet the timing easily without the need for buffer insertion. Since the neurons are pin-capacitance and internal-power dominant designs, reducing the buffer count in M3D design plays an important role in the power savings.

6.3.4 Full-Chip Results

Figure 6.6 shows how the smaller individual neuron enabled by architectural optimization impacts on the full-chip footprint, wirelength and static power consumption. Compared to

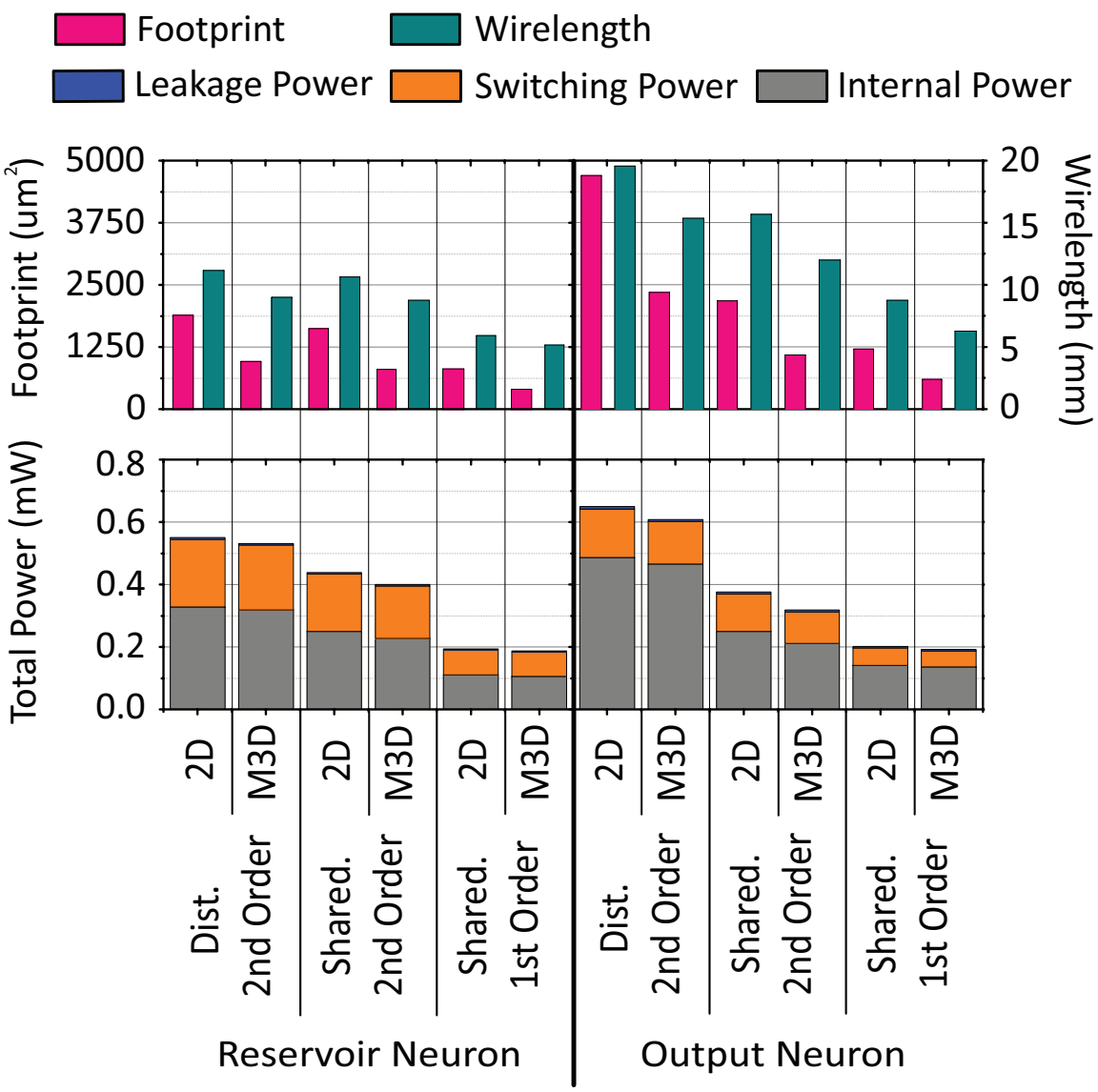


Figure 6.5: Individual 2D and M3D neuron implementation results used to build full-chip LSM neuromorphic processor with the architectural combinations based on the proposed memory sharing and controlling the synapse model complexity.

the baseline architecture, full-chip footprint of the shared memory 2nd-order and 1st-order architecture is reduced by 21% and 53%, respectively while keeping the same spacing between the neuron blocks at the top-level placement. However, in shared memory 2nd-order architecture, we observe that this footprint savings does not lead to the wirelength savings because of the routing overhead from the shared memory to the individual neurons. Instead, the shared memory helps to reduce the full-chip internal power by 23%, and this leads to

18% of total power savings. On the other hand, shared memory first-order architecture has both wirelength and power savings by 35% and 55%, respectively.

At the top-level, M3D ICs have clear wirelength savings from the 2D counter parts at the same architecture thanks to the large number of inter-neuron connectivities. In every architecture, M3D designs offer more than 24% inter-neuron wirelength savings. However, we observe that this inter-neuron wirelength savings do not guarantee the huge full-chip switching power savings because of the sparse communications between the neurons in the LSM processer. Nonetheless, combining all the power savings from both individual neurons and the top-level, we find that both architectural optimization approaches help to increase the M3D power savings from 9% to 13%.

6.4 Application-based Analysis

We carry out the real-world application of speech recognition on the implemented LSM neural processors and the practical 3D IC benefits. The benchmark is adopted from the TI46 speech corpus [61], which contains read utterances from 16 speakers of the English letters 'A' through 'Z'. Without loss of generality, we select one representative speech for the letter 'R' and evaluate the power dissipation in our designs. The continuous temporal samples are preprocessed by Lyon's ear model [62] and encoded into 78-channel spike trains using the BSA algorithm [63]. These speech patterns are sent to both 2D IC and 3D IC designs. The labeled 26 output neurons correspond to the 26 letters in the English alphabet and the output spike trains of the intended output neuron ('R' in this case) is observed as expected.

6.4.1 Full-Chip Power Breakdown

Figure 6.7 shows the power consumption results for the reservoir and output training, and classification of the letter 'R' from three-different architecture presented in this work. Thanks to the clock gating implementation, the different activation of training and reservoir

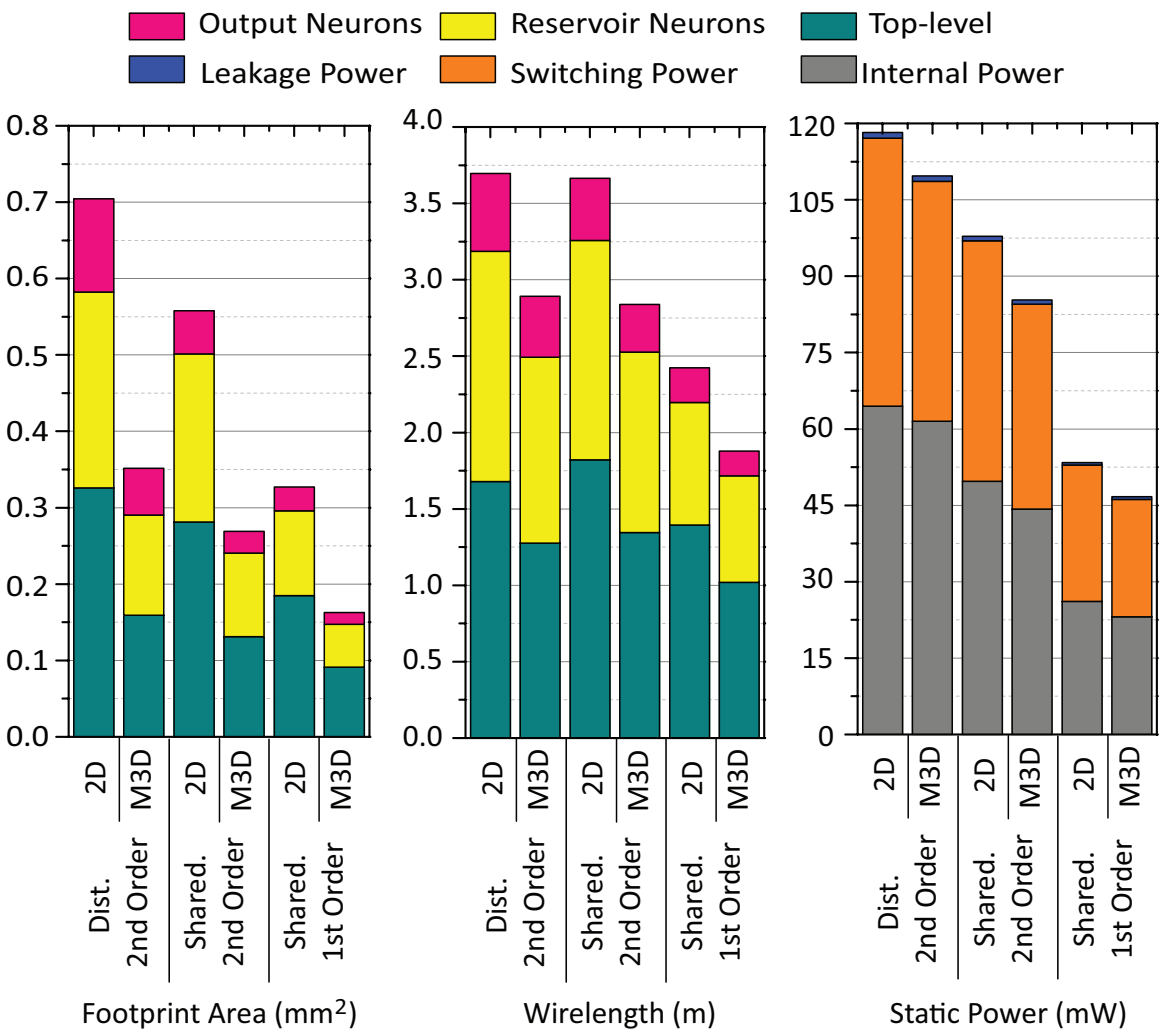


Figure 6.6: The impact of shared memory and synaptic models on the full-chip design results.

unit effectively reduces the total power consumption. In the reservoir training phase, there is no power consumption of the training unit (output layer) as its clock is completely gated out. During the output training and testing phases, the power of reservoir unit is much smaller than the reservoir training phase because reservoir synaptic weights do not change. Architectural optimization has a great impact on the total power savings. Compared to 2D ICs with distributed memory, 2D shared memory design with second- and first-order architecture offer 36% and 57% power savings for reservoir training, and 4% and 27% for output training, respectively.

For the testing, we observe 7% and 38% of power savings, respectively. The major source of these huge power savings are derived from the individual reservoir neuron optimization. Regarding the M3D power savings, we find M3D designs always reduce the top-level power consumption by more than 20%. However, as a part of the overall bio-inspired computation models, the recurrent SNN inherently operated with sparse firing activities, therefore power savings at the top-level inter-neuron communications have been generally consistent and small. Another benefit from M3D is the output neuron power savings. We observe that output units have a maximum of 12% power savings in M3D compared to the 2D counterpart, and this leads to clear power savings in M3D for output training and actual classification.

6.4.2 Power-Performance-Area-Accuracy Benefit

The energy dissipation is dependent on the power as well as the number of clock cycles of operation. Although the shared memory architecture offers huge footprint and power savings, the shared reservoir memory requires additional clock latency to access compared to the flip-flops in the distributed reservoir weight storage. The design with first-order synaptic model also largely saves power and footprint, but this hurts the classification accuracy from 92.3% to 91.9%. Therefore, we compare the final power-performance-area-accuracy benefit of the design and architectural co-optimization in LSM neuromorphic processor to measure the tradeoff among different design criteria.

Although different input letters propagate different firing activities in the system, the input spike activity only determines the top level activity, which is a very small part of the total energy of the system.

In this work, we calculate the average energy consumption for training and classifying a representative speech sample. In general, the overall spike density is roughly the same over various samples. Therefore, the average power remains the same and we use the power consumption values for each phase from Section 6.4.1. To get good learning performance

Figure 6.7: Vector-based power consumption analysis in different operation steps

over the entire benchmark, 25 epochs of reservoir training and 250 epochs of output training are conducted and these numbers of iterations are taken into account when calculating the total energy consumption.

Targeting 1GHz clock operation, Table 6.1 summarizes the overall energy savings for 2D ICs and 3D IC LSM neuromorphic processor based on the three-different architecture, with two-different design approaches, respectively. Although the reservoir training energy is actually large in shared memory architecture, it has little impact on the total energy dissipation considering its small number of training iterations than the output training. Also, the power and footprint savings are significantly large over the accuracy degradation when

using first-order synaptic model. This implies that the power and footprint savings from our co-optimization approaches are well preserved in the energy consumptions for the speech recognition. On average, for the LSM neural processor, M3D IC design gives up to 19% less energy consumption than its 2D IC counterparts for training and inference of a speech sample. Overall, we observe 70% power-performance-area-accuracy benefit from using design and architectural co-optimization compared to the 2D baseline design.

Table 6.1: Power $\times$ Operation Time Period $\times$ Silicon Area $\div$ Accuracy (PPAA) benefit of design and architectural co-optimization proposed in this work.

	Distributed Second-order		Shared Second-order		Shared First-order	
	2D	M3D	2D	M3D	2D	M3D
Silicon Area (mm^2)	0.070	0.070	0.056	0.054	0.033	0.033
Res. Tr. Period (ms)	1.35		3.42			
Res. Tr. Power(mW)	87.76	76.93	56.39	53.68	37.84	35.32
Res. Tr. Energy(mJ)	118.88	104.21	192.77	183.51	129.36	120.74
Out. Tr. Period (ms)	109.40		109.41			
Out. Tr. Power(mW)	35.92	33.70	34.46	28.70	26.17	23.28
Out. Tr. Energy(mJ)	3.929	3.687	3.770	3.140	2.863	2.547
Training Energy (mJ)	4.048	3.791	3.963	3.323	2.993	2.668
Test Period (ms)	0.21		0.24			
Test Power (mW)	46.37	41.85	43.22	36.92	28.85	26.05
Testing Energy (mJ)	0.009	0.008	0.010	0.009	0.007	0.006
Total Energy (mJ)	4.058	3.799	3.973	3.333	2.999	2.674
Accuracy (%)	92.3				91.9	
Normalized PPAA	1	0.93	0.77	0.62	0.34	0.30

6.5 Conclusion

In this work, we implemented M3D IC design for an LSM-based neuromorphic processor and devised various design and architectural co-optimizations to minimize the energy consumption in the speech recognition. We presented the impact of shared memory architecture and the impact of the synaptic model complexity on the individual neuron and full-chip design. We measured the energy dissipation for speech recognition application with TI46 corpus spoken English speech samples, and achieved up to 70.0% reduction in the power-performance-area-accuracy overhead. This work serves as an important step

towards realizing bio-inspired neuromorphic processors utilizing 3D IC design advantages.

CHAPTER 7

AREA-EFFICIENT AND LOW-POWER GATE-LEVEL FACE-TO-FACE-BONDED 3D LIQUID STATE MACHINE DESIGN

In this work, we adopt face-to-face(F2F)-bonded 3D integration technology to the LSM neuromorphic processor designs. The advancement of F2F wafer-level bonding technology enabled the bonding precision less than $0.5\mu m$ [64]. This allows us to enable fine-grained 3D interconnections to maximize 3D IC benefits. Using the state-of-the-art RTL-to-GDSII physical design flow name Compact-2D [65], we explore the power-area-accuracy benefits of F2F-bonded gate-level 3D LSM processors targeting the next generation neuromorphic processors. The major contributions of this work are as follows:

- We study how the different size of a reservoir in the LSM architecture affects the learning performance, power consumption, and design area.
- We analyze the impact of reservoir connectivity density on power-area-accuracy trade-off in LSM processor designs.
- We design a commercial-quality F2F-bonded 3D LSM IC with the optimal LSM architecture and explore the 3D integration benefits in the LSM processor designs.

7.1 Liquid State Machine

7.1.1 System Architecture

The targeted LSM system architecture is adopted from Chapter 6. It consists of a reservoir stage and a training stage, where a number of digital reservoir neurons and readout neurons are instantiated respectively. External spike inputs to the reservoir stage are assigned to their targeted neurons through a pre-defined crossbar interface, and the spikes generated

from the reservoir stage are sent to all readout neurons and also back to some other reservoir neurons through a crossbar interface.

One thing to mention is that, in this work, we adopt the idea of using a large stage-wised weight storage memory for all readout neurons to reduce the large peripheral overhead when instantiating individual memory inside each neuron proposed in Chapter 6. The memory sharing mechanism is based upon the property of the proposed LSM design that all readout neurons work in parallel and the synaptic weights are accessed in serial following the same order. Therefore, all neurons tend to read and write the same memory address at every emulation time step. We then combine the weights previously stored at the same address in the individual memories in a row at the same address of the shared memory. When accessing the memory data, the weights of different neurons are divide assigned to corresponding neurons when reading or concatenated from all neurons for when writing.

7.1.2 Training Algorithms

The training of the LSM processor is executed in two stages. First, the reservoir stage is trained by the a hardware-friendly spiking dependent plasticity (STDP) algorithm adopted from [59] until the synaptic weight distribution converges. STDP is an unsupervised Hebbian learning mechanism updating synaptic weights based on the temporal relationship of pre- and postsynaptic spikes:

$$\begin{aligned} \Delta w^{+} &= A_{+}(w) \cdot e^{-\frac{|\Delta t|}{\tau_{+}}} \quad if\, \Delta t > 0 \\ \Delta w^{-} &= A_{-}(w) \cdot e^{-\frac{|\Delta t|}{\tau_{-}}} \quad if\, \Delta t < 0, \end{aligned} \tag{7.1}$$

where Δw^{+} and Δw^{-} are the weight modifications induced by long-term potentiation (LTP) and long-term depression (LTD), and $A_{\pm}(w)$ determines the strength of LTP/LTD.

Implementing the STDP learning rule accurately on hardware produces good performance, however, at a cost of high discretized bit resolution and frequent weight update which leading to large area/power overhead. However, employing it with aggressively dis-

cretized synaptic weights and the STDP learning curve leads to an immediate performance degradation. To address this problem, the adopted hardware-friendly STDP is realized by a look-up table based implementation with minimal aggregated discretization error and simple logic.

Second, during the readout training phase, a biologically plausible supervised spike-based algorithm [25] is employed to perform the classification. To introduce the supervision on spikes, an additional stimulus is injected into each readout neuron to bring up the firing activity to the desired level, which is characterized biologically by a Calcium concentration model. The Calcium concentration C is computed by:

$$C(t) = C(t-1) - \frac{C(t-1)}{\tau_c} + S(t), \tag{7.2}$$

where $S(t)$ is the spiking event at the current time step. Then, the readout synapse weight is update statistically:

$$\begin{aligned} w_i &= w_i + \Delta w \;\; with\; P_+, \quad if \;\; C_\theta < C < C_\theta + \Delta C \\ w_i &= w_i - \Delta w \;\; with\; P_-, \quad if \;\; C_\theta - \Delta C < C < C_\theta, \end{aligned} \tag{7.3}$$

where P_+ and P_- are the probabilities for potentiation and depression, and C_θ and ΔC are the calcium concentration threshold and margin width respectively. Moreover, during the readout training phase, reservoir stage continues to be activated to provide spike inputs to training stage while maintaining its synaptic weights.

7.2 Design and Simulation Setting

In this work, we adopt the Compact-2D flow proposed in Chapter 5 to build F2F-bonded gate-level 3D LSM processor. In order to thoroughly assess the learning performance and power-area benefits of F2F-bonded 3D LSM processor architecture, we first present the impact of reservoir size and the connectivity density on the learning performance and the

power-area overhead in 2D LSM processors. Then, we analyze the power-area-accuracy benefits from F2F-bonded 3D LSM processors compared with 2D LSM designs.

7.2.1 LSM Design Generation

Four LSM designs with different reservoir sizes are generated to study the impact of reservoir size. Each external input spike is connected to 8 randomly reservoir neurons with a fixed weight, and this fixed weight is randomly chosen to be 8 or -8 with equal probability. For internal connections in the reservoir stage, we follow the widely used settings suggested in [66], which is based on the microcircuit in the real biological brain in that 80% of the reservoir neurons are excitatory and 20% of the reservoir neurons are inhibitory. Synaptic connections are initialized stochastically according to the Euclidean distance between pre- and post-synaptic neurons. The probability of creating a connection between the neuron u and v is calculated by:

$$p = C \cdot exp\left[-(\frac{D(u,v)}{\lambda})^2\right]. \tag{7.4}$$

where λ is a connection parameter, and $D(u,v)$ is the Euclidean distance between the neuron u and v.

To study how the connectivity density affects the performance and the hardware costs of the LSM processor, we set the two representative values of parameter C for each network size, which are 1.5 and 4.5. As a result, the design with $C = 4.5$ has roughly three times more reservoir synapses than the design with $C = 1.5$. In our LSM hardware implementation, all reservoir neurons process information in parallel and also in a synchronous manner. However, inside each neuron, the pre-synaptic spike inputs are examined in serial in the synaptic input processing module as mentioned in Section 7.1. Therefore, even though each reservoir neuron may have different number of synaptic connections, the number of registers that stores the variables corresponding to the synapses is the same among all reservoir neurons in a single LSM design and is decided by the actual maximum reservoir

Table 7.1: Key components in a reservoir neuron of our LSM designs. RVN72 denotes a design with 72 reservoir neurons in the reservoir stage, etc.

LSM Designs	RVN72	RVN90	RVN112	RVN135
Basedline Design: $C = 1.5$ in Equation (7.4)				
Shift Register	22	18	20	16
Input Synapse Width	4	4	5	5
Reservoir Synapse Width	5	5	5	4
Input Spike	16	16	21	32
Reservoir Spike	22	18	20	16
Dense Reservoir Design: $C = 4.5$ in Equation (7.4)				
Shift Register	31	38	48	35
Input Synapse Width	4	4	5	5
Reservoir Synapse Width	5	5	6	6
Input Spike	16	16	21	32
Reservoir Spike	31	38	48	35

synapse connections among all reservoir neurons. However, as introduced in Equation 7.4, the synaptic connection between two neurons is a random variable, it is possible that a network actually instantiates more resources to implement the synapses although it has a smaller reservoir size. Table 7.1 tabulates the list of key components in a reservoir neuron from the four different benchmarks.

7.2.2 LSM Performance Simulation Setting

To measure the learning performance of proposed neural processors, we choose a representative non-trivial real-world benchmark of speech recognition, which is a subset of the TI46 speech corpus [61]. The benchmark has 260 speech samples of ten spoken utterances of English letters from "A" to "Z". The continuous temporal speech signals are preprocessed by Lyon's ear model [62] and then encoded into 78 spike trains using the Bens Spiker Algorithm [63]. We adopt a 5-fold cross validation scheme in the standard machine learning process to assess the learning performance.

7.3 2D IC Design Results

7.3.1 Impact of Reservoir Size

For the full-chip 2D LSM designs, we use four metal layers in a commercial-grade 28nm Process-Design-Kit, and use different sizes of register file for the shared memory in the readout stage depending on the reservoir size. The maximum target clock frequency is 1.3GHz for all the designs. Figure 7.1 shows the silicon area and the interconnections colored by the characteristics of connectivity in four different 2D LSM designs. The number of reservoir neurons in the reservoir stage is used to name those designs. The RVN135 LSM design requires 78% more silicon area, and 101% more total wirelength compared with those of RVN72 for the 2D full-chip implementation. Figure 7.2 shows the detailed wirelength distribution divided by the purpose of the interconnection. As the reservoir size increases, not only the wirelength for the intra-RVN is increased in proportion to the reservoir population, but also the inter-RVN wirelength is increased since the reservoir neurons are spread out of the entire design. This contributes to additional wirelength for full-chip implementation. Compared to RVN72, the inter-RVN wirelength for the RVN135 design is exponentially increased by 5.7x while intra-RVN wirelength is increased by 2.6x. In addition, it is worth noting that RVN112 has only 1.8% less total wirelength while the silicon area is 10.3% smaller than RVN135. This is because RVN112 has inherently larger reservoir spike connectivity than the other designs as presented in Table 7.1. It turns out that RVN112 actually has 19.3% more wirelength for inter-reservoir neuron connections than RVN135 design. This indicates that the fine-tuned parameter optimization for the initial network generation is critical for the area-efficient low-power LSM designs targeting the edge-computing devices.

To summarize the impact of reservoir size on the power-area-accuracy of 2D LSM designs, Table 7.2 tabulates the detailed design and performance metrics. The target clock frequency is 1.3GHz, and the static power analysis is performed based on the 0.1 switch-

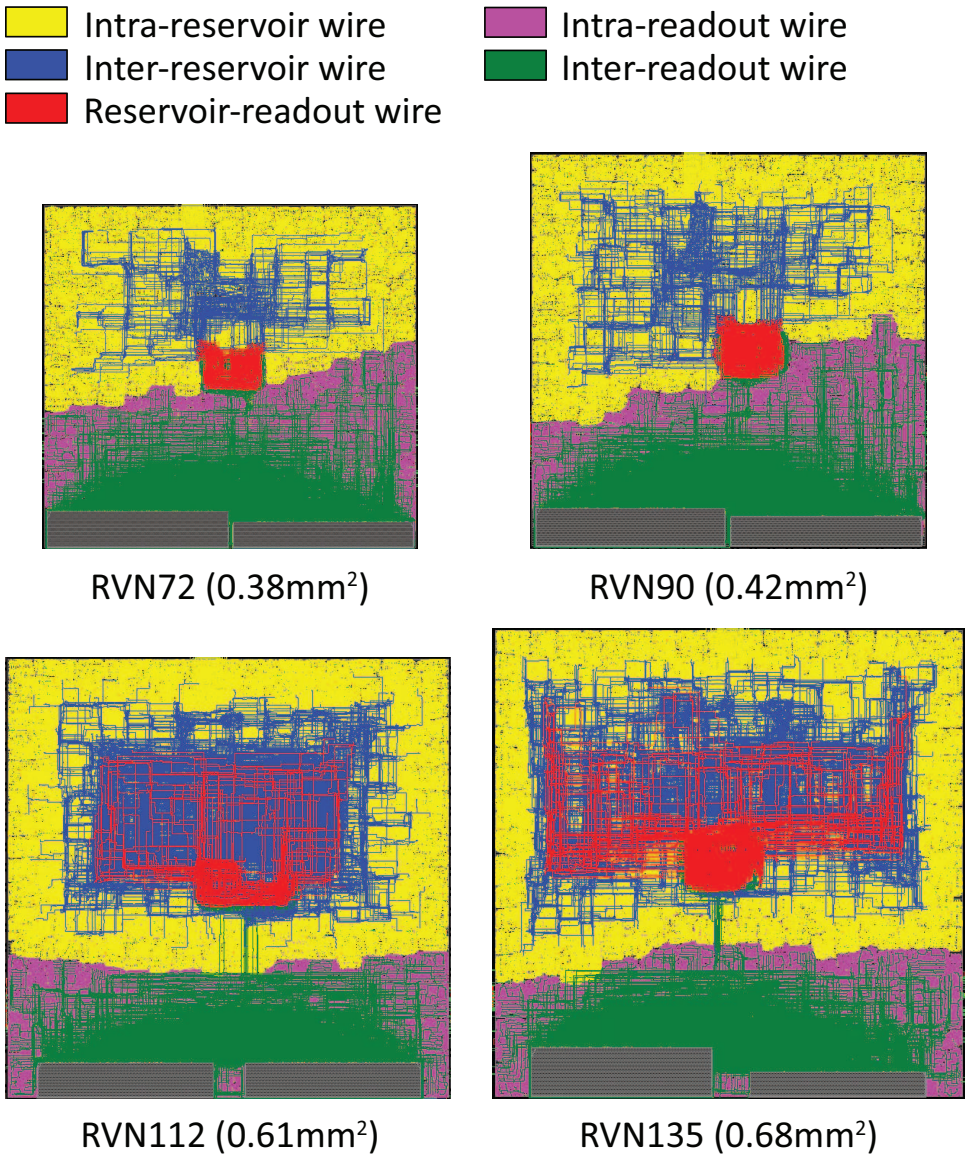


Figure 7.1: 2D full-chip LSM designs. Larger reservoir size increases the design footprint significantly.

ing activity for the primary inputs and D-Flip Flop outputs, and 2.0 for the clock. The placement utilization of all designs is targeted within 75% 78% range for the fair design comparison. We first observe that larger reservoir network size gives us the better classification accuracy. RVN135 design achieves 92% speech recognition accuracy while RVN72 design has only 88% classification performance. This makes sense since a reservoir with more neurons holds a more complex non-linear dynamics, and can map the input

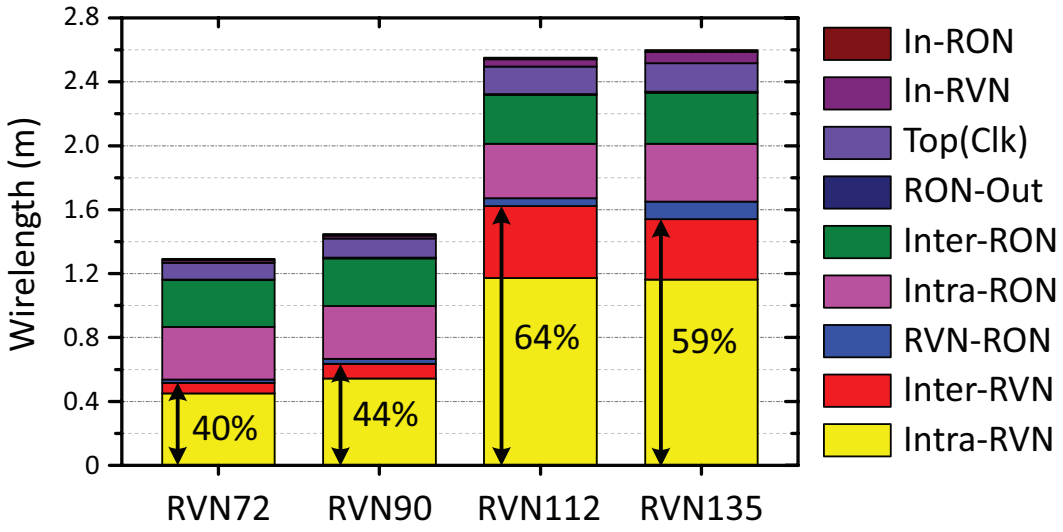


Figure 7.2: Wirelength distribution of the 2D LSM designs in Figure 7.1. RVN denotes reservoir neurons, and RON readout neurons. As the reservoir size increases, wirelength from the reservoir network becomes dominant while the others remains relatively the same.

patterns to a higher-dimensional feature space which provide rich information for the readout layer for classification. However, this improvement is not for free with regard to the silicon area, and total power consumption. Compared with RVN72, 78% of form factor increase in RVN135 leads to more timing buffers on top of the increased wirelength, and they contribute to 96.4% more total power consumption in RVN135.

7.3.2 Impact of Reservoir Connectivity

Figure 7.3 shows the impact of reservoir connectivity density on the wirelength for the intra- and inter- reservoir neurons in four LSM designs with different network size. We observe that the increased reservoir connectivity affects the intra-reservoir neuron wirelength. This is because each reservoir neuron contains more shift registers and reservoir spike storages in its implementation to support the dense reservoir connecitivity as presented in the Table 7.1. The intra-reservoir neuron wirelength is increased by 67% in average. Regarding the inter-reservoir neuron wirelength, we observe 3.69x more wirelength in the dense reservoir designs in average. A maximum 5.42x increase is observed in RVN90 design, and a minimum 1.84x increase in RVN112 due to RVN112's inherent enhanced reservoir

Table 7.2: 2D LSM designs with coarser reservoir connectivity. The target clock frequency is 1.3GHz. RVN135 design shows 1.78x more silicon area, and 1.96x more power consumptions compared with RVN72 design, while improving the classification accuracy.

Baseline Neuron Connectivity				
2D LSM Designs	RVN72	RVN90	RVN112	RVN135
Worst Neg. Slack (ps)	1.28	1.60	13.01	12.72
Liquid Neuron Count	72	90	112	135
Classification Accuracy	87.69%	88.85%	90.76%	91.54%
Placement Util. (%)	75.10	76.50	78.02	77.63
Silicon Area (mm^2)	0.378	0.421	0.613	0.678
Total Wirelength (m)	1.292	1.447	2.552	2.598
Seq. Cell Count	6,893	7,966	14,406	15,572
Comb. Cell Count	81,043	91,284	135,871	150,517
Total Cell Count	87,936	99,250	150,277	166,089
Pin Capacitance (pF)	226.2	261.1	393.9	436.7
Wire Capacitance (pF)	112.7	126.2	234.0	232.6
Total Capacitance (pF)	338.9	387.3	627.9	669.3
Wire Cap. Ratio (%)	33.25	32.58	37.27	34.75
Switching Pwr (mW)	46.88	54.03	86.62	90.92
Internal Pwr (mW)	21.43	24.86	39.38	42.74
Leakage Pwr (mW)	6.25	7.26	11.64	12.75
Total Pwr (mW)	74.56	86.15	137.64	146.41

connectivity density decided by the random process when we generate the design.

To analyze the power-area-accuracy impact of the reservoir connectivity density, Table 7.3 tabulates the performance and design metrics of dense reservoir designs. For the fair comparison, the metal layer usage, maximum target clock frequency, and the static power analysis setting is the same with the baseline designs. First, we observe increased reservoir connectivity density gives us additional classification accuracy improvement without increasing the reservoir size. when the recurrent reservoir connection gets denser, the diversity of reservoir dynamics is improved with a better interaction among input samples, thus boost the learning performance. However, if the recurrent connections in a reservoir is too dense, there would be a chaos inside the reservoir and no performance improvement will be observed. Therefore we do not see an obvious performance improvement for larger network size in which the overall synaptic connections are already large. Although this

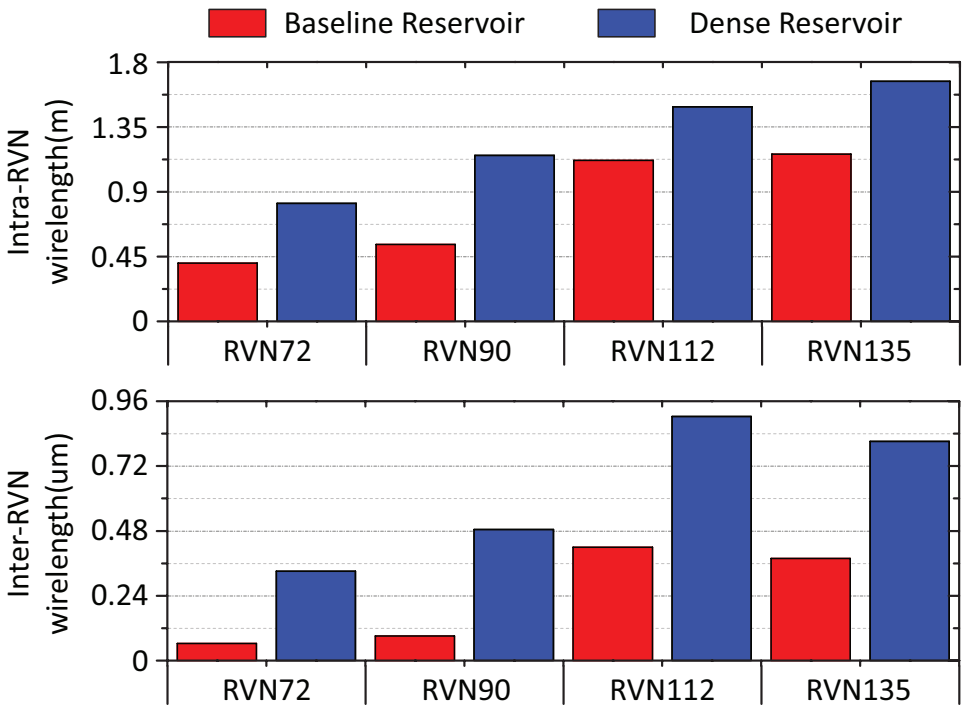


Figure 7.3: Impact of reservoir connectivity on the inter-RVN, and intra-RVN wirelength of 2D LSM designs. We observe 1.67x intra-RVN wirelength increase and 3.69x inter-RVN wirelength increase in the designs with dense reservoir.

small accuracy improvement is significant and impressive in the speech recognition field and also for the edge-computing devices, it turns out the area-power expense to enable the dense reservoir connectivity is not ignorable. We find that 1.28x more silicon area and 1.38x more power consumptions is required in average for the designs with denser reservoir connectivity compared with their baseline. This indicates that if we put the accuracy to the highest priority on the LSM design, larger reservoir network with reasonable reservoir connectivity offers better power-area-accuracy tradeoffs in LSM design implementation.

7.4 3D IC Design Results

To preserve the accuracy benefit while minimizing the form factor and power consumption, we build two-tier F2F-bonded 3D LSM designs. For the experiments, we focus on the RVN135 architecture with baseline reservoir connectivity, which gives the maximum classification performance among the different reservoir sizes, and better area-power results

Table 7.3: 2D LSM designs with denser reservoir connectivity. Δ denotes increase compared with the baseline connectivity shows in Table 7.2. We observe slight increase in accuracy, 1.28x more silicon area and 1.38x more power consumption.

Dense Neuron Connectivity				
2D LSM Designs	RVN72	RVN90	RVN112	RVN135
Worst Neg. Slack (ps)	-11.98	-9.94	-14.65	-11.77
Classification Accuracy	88.46%	89.23%	90.77%	91.54%
Placement Util. (%)	77.86	77.65	78.39	78.20
Silicon Area (mm^2)	0.492	0.596	0.757	0.814
Total Wirelength (m)	2.019	2.490	3.329	3.526
Seq. Cell Count #	11,709	15,292	20,679	20,919
Comb. Cell Count #	110,424	129,489	164,456	180,230
Total Cell Count #	122,133	144,781	185,135	201,149
Switching Pwr (mW)	66.51	83.00	108.87	116.69
Internal Pwr (mW)	30.67	39.04	51.21	54.16
Leakage Pwr (mW)	9.17	11.43	15.30	16.22
Total Pwr (mW)	106.36	133.47	175.38	187.07
Total Pwr Δ	1.4x	1.5x	1.3x	1.3x

compared with the dense reservoir connectivity. For the full-chip two-tier F2F-bonded 3D LSM designs, four metal layers are used for each die. F2F via size is assumed to be 0.5um, the pitch is 1.0um, resistance is 0.2ohm and 0.5fF for via capacitance. For the initial floorplan, two register files are split into different tiers and vertically overlapped to minimize the form factor occupied by them. Figure 7.4 shows GDS layouts.

In Table 7.4, we present 2D vs. F2F-bonded 3D RVN135 designs at iso-performance. For fair comparison, placement utilization for 2D and both dies of F2F design is set to be the same. First, we achieve more than 50% footprint saving with F2F compared with 2D. This leads to 4% silicon area saving under comparable placement density (78.63% vs. 79.06%). We believe this is significant as most of the work published in the literature show 50% footprint saving, which means zero silicon area saving. Moreover, we achieved this area saving under the identical classification accuracy. We believe such a footpring saving is useful for Internet-of-Things sensors that require smallest possible form factor. The silicon area saving directly impacts cost.

The wirelegnth (2.598 vs. 2.119) and cell count saving (160K vs. 166K) in F2F leads

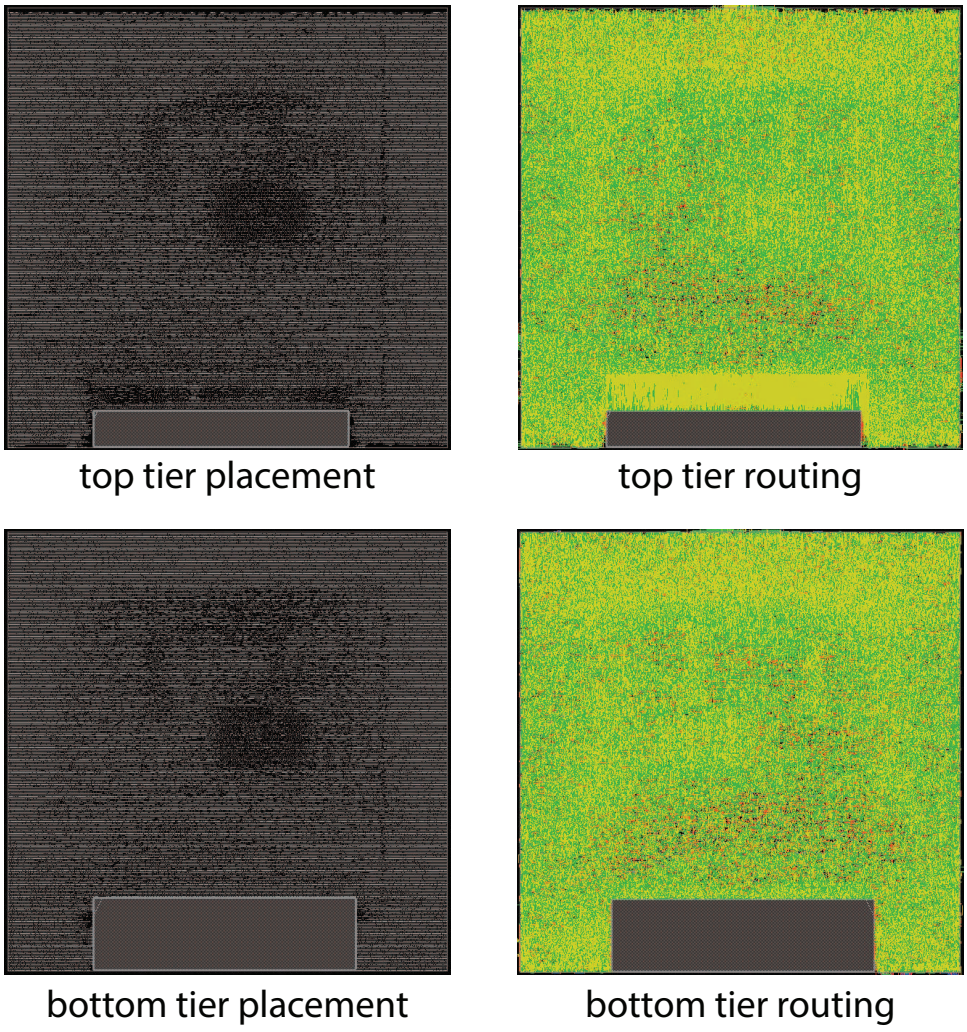


Figure 7.4: Face-to-face two-tier 3D IC layout of our RVN135 architecture with baseline reservoir connectivity.

to 3% total power saving. Power saving is not as significant as our LSM architecture is pin-cap dominated: in 2D design, pin:wire capacitance ratio is 436.7: 232.6. In 3D design, 423.5: 199.2. The wirelegnth savign in 3D design only affects wire capacitance, thus the small power saving.

Another reason for small power saving is due to the cell displacement introduced during Compact-2D flow. In Compact-2D, cell overlap after tier partitioning is removed using a legalizer, which necessitates cell displacement in both tiers. Figure 7.5 shows the difference between the X and Y cell location after tier partitioning and those of the final F2F design. while 34.8% cells keep the optimal placement location, 63.6% of cells change their location

Table 7.4: F2F-bonded 3D RVN135 vs. 2D RVN135. F2F achieves 52% form factor savings, 4% silicon area savings, and 3% total power savings under the same 92% accuracy.

Designs	2D	F2F
Worse Neg. Slack (ps)	-12.72	-9.53
Form Factor (mm^2)	0.678	0.326
Silicon Area (mm^2)	0.678	0.652
Placement Util. (%)	78.63	79.06
Total Wirelength (m)	2.598	2.119
Seq. Cell Count	15,572	
Comb. Cell Count	150,517	144,922
Total Cell Count	166,089	160,494
Pin Capacitance (pF)	436.7	423.5
Wire Capacitance (pF)	232.6	199.2
Total Capacitance (pF)	669.3	622.7
Wire Cap Ratio (%)	34.95	31.99
Switching Power (mW)	90.92	86.45
Internal Power (mW)	42.74	42.34
Leakage Power (mW)	12.75	12.91
Total Power (mW)	146.41	141.90
Accuracy	92%	

from $1\mu m$ to $4\mu m$. 0.19% of cells more than $10\mu m$ displacement while the post-tier-partitioning optimization. The total displacement is $0.21m$, and this partially accounts the wirelength increase (and power increase) in the final F2F design.

Figure 7.6 presents wirelength distribution comparisons. The wirelength of Compact-2D nets is scaled by 0.693 to show the difference between the electrical length of the Compact-2D nets and the actual wirelength of F2F nets. Most of small nets less than $20\mu m$ has negligible impact on the final F2F nets, but 1% of these small nets are affected by 3D routing and the resulting wirelength become larger than $20\mu m$. This is not considered during the Compact-2D design, and the nets between $20\mu m$ and $40\mu m$ has increased by 18%. Along with the placement difference, this 3D routing causes wirelength increase in F2F design, and looses the wire capacitance savings, and switching power savings obtained in the Compact-2D design. The final total power savings of F2F design turns into the 3%, with the significant form factor savings by 48%.

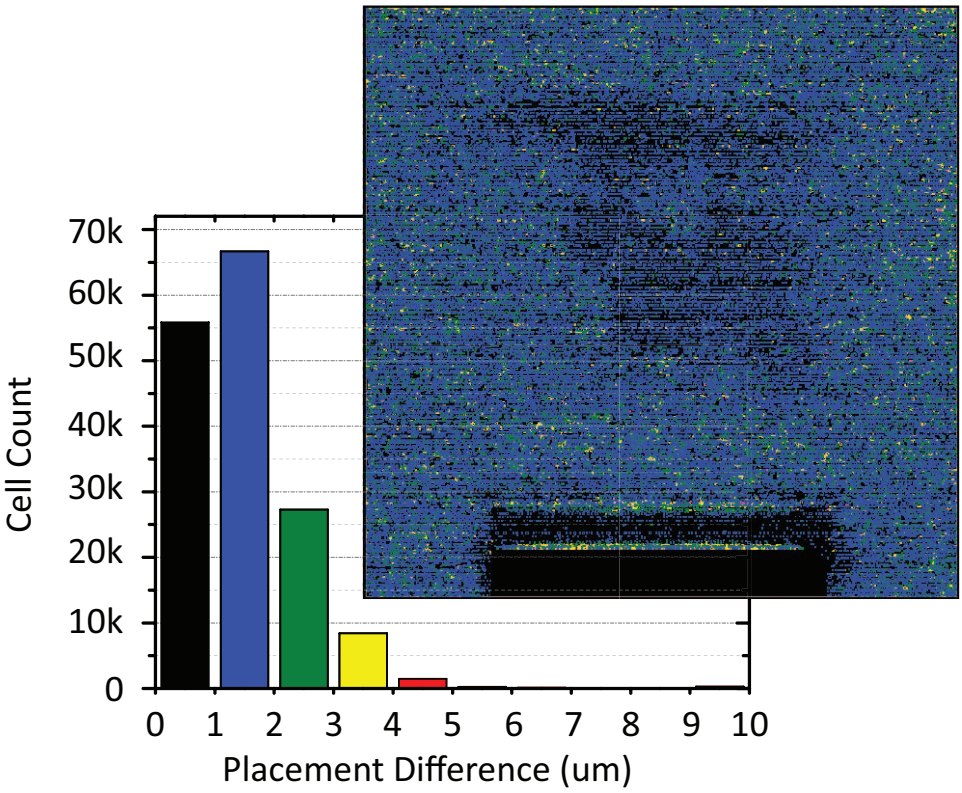


Figure 7.5: Cell displacement before and after tier partitioning. Cells are moved to remove the overlaps caused by the placement contraction in Compact-2D [65]. We observe 65.2% of the cells change their location. The total displacement is $0.21m$. The yellow cells in the die shot show displacement.

7.5 Conclusion

In this work, we studied the impact of the reservoir size and the connectivity density on the classification accuracy and their area-power overhead of the liquid state machine (LSM) processor designs. We showed that the 135 reservoir neurons, which is the maximum size of the reservoir used in this work, significantly improves the learning performance of a representative non-trivial real-world benchmark of speech recognition by 4.4% compared with the design with 72 reservoir neurons. However, we observed that this accuracy improvement is not for free and requires 78% more form factor, and 96.4% power consumption. Regarding the reservoir connectivity, we presented that denser reservoir improves the accuracy by a maximum 0.8% without increasing the network size, but this results in 28%

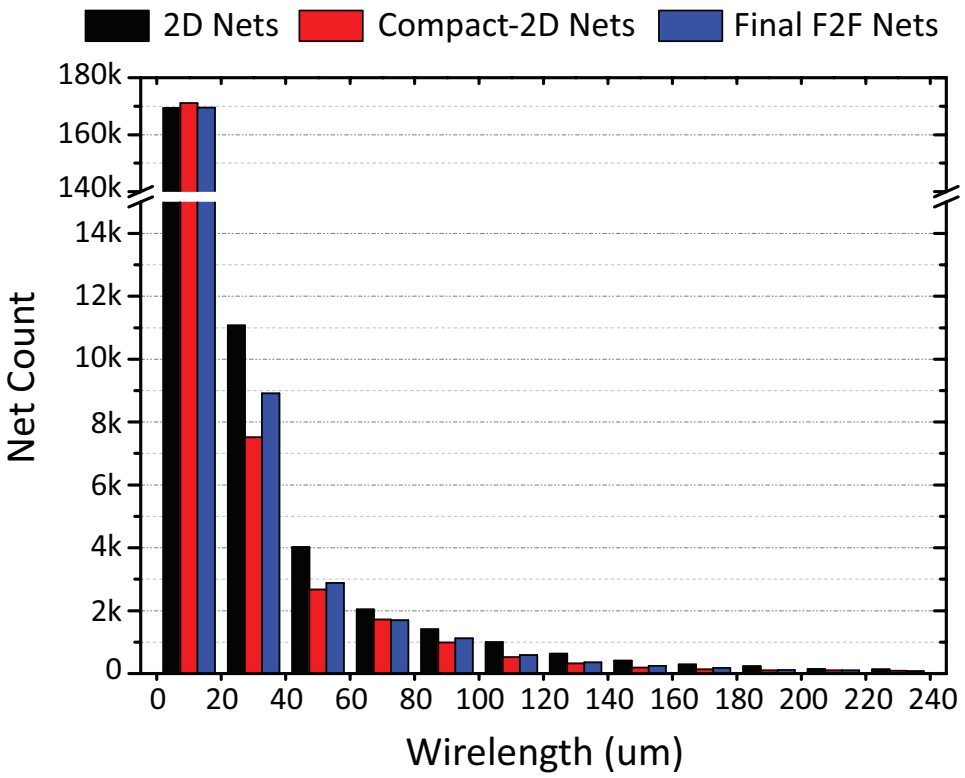


Figure 7.6: Wirelength distribution comparison among 2D, Compact-2D, and final F2F-bonded 3D LSM designs.

more form factor, and 38% more power consumption in average.

Lastly, we explored the F2F-bonded 3D integration benefits on the LSM processor design using the state-of-the-art physical design flow name Compact-2D, and thoroughly analyzed the area-power benefits in the two-tier F2F-bonded 3D LSM processor design. We observed that F2F-bonded 3D integration has significant benefits on the form factor savings, and additional power savings while preserving the great classification accuracy. Using the design with 135 reservoir neurons, the F2F-bonded 3D LSM processor achieved 52% form factor savings, which is smaller than the 2D design with 72 liquid neurons by 13.8%, and additional 3% power savings compared to its 2D counterpart. This work suggests the power-area-accuracy tradeoffs on the reservoir optimization in the 2D and F2F-bonded 3D LSM processors targeting the next generation neuromorphic edge-computing devices in the Internet-of-Things Era.

CHAPTER 8

SUMMARY AND FUTURE DIRECTIONS

8.1 Summary and Conclusions

8.1.1 T-M3D Standard Cell Layout Optimization for Full-Chip Static Power Integrity

In this study, we proposed a new layout optimization method, the stitching scheme, for the transistor-level monolithic 3D (T-M3D) standard cell design. The stitching scheme addresses the static power integrity issue inherent in the folding scheme for T-M3D cell layouts. It also minimizes the timing/power degradation caused by parasitics originating from the unique T-M3D layout architecture. We developed the 14*nm* T-M3D technology process design kit and designed 41 standard cells in the form of 2D, folding T-M3D, and stitching T-M3D layouts. We proved that the stitching scheme outperforms the folding scheme in terms of timing and power metrics at the expense of the increase in the cell height by only 0.5 metal tracks. We also presented a design methodology for a power delivery network in folding T-M3D ICs, and performed sign-off IR-drop analysis in both folding and stitching T-M3D ICs. Lastly, we found that the folding scheme cannot be applied to commercial grade layouts because of its severe IR-drop. However, compared to 2D ICs, the stitching T-M3D ICs experience only 6mV increase in maximum IR-drop while reducing the footprint by up to 44% and power consumption by 6%.

8.1.2 Cost Overhead to Justify the Adoption of Monolithic 3D IC at 7nm Era

This study showed power, performance, and cost (PPC) tradeoffs with full-chip GDS based cost modeling for 2-tier, gate-level, full-chip GDS M3D ICs built using a foundry-grade 7nm bulk FinFET technology. We proposed normalized wafer and die cost models based on the number of metal stacks and die area for 2D and M3D. In our PPC tradeoff study with the

simple but self-contained cost models, both 2D and M3D designs are optimized in terms of the number of BEOL metal layers used for routing to obtain the best possible PPC values for the fair comparison. Also, a new CAD methodology for 2-tier G-M3D named Projected-2D Flow is developed. Projected-2D maximizes the placement and routing utilization of an M3D design by reducing its footprint by more than 50% compared with that of the 2D counterpart. Furthermore, this flow allows us to accurately compare RC parasitics of equivalent nets in both 2D and M3D designs since final netlists of these two design flavors are the same.

Based on the experiments with two widely different circuit types (BEOL-dominant vs. FEOL-dominant), it is confirmed that while M3D has indeed a great footprint saving, the PPC quality of M3D is actually worse than that of optimized 2D reference by 34% due to high M3D wafer cost. Our study also showed that, for the adoption of M3D technology at the 7nm era, M3D wafer yield needs to be higher than 90% of 2D wafer yield, and the 2-tier device manufacturing cost of M3D design needs to be limited by less than 33% of 2D device manufacturing cost, and lastly the die area should be large enough ($100mm^2$-scale) to have fruitful die cost reduction from huge M3D footprint saving. Lastly, and counter-intuitively, this study showed that FEOL-dominant type circuit has PPC benefits from M3D technology more and sooner than BEOL-dominant type circuit.

8.1.3 Physical Design Solutions to Tackle FEOL/BEOL Degradation in G-M3D ICs

In this research, we proposed CAD methodologies for gate-level monolithic 3D ICs (M3D) that tackle the FEOL/BEOL inter-tier variations caused by low temperature manufacturing. To address the top tier device degradation, we presented a cell-slack sorting-based tier partitioning algorithm that assigns timing critical elements into the bottom tier. To deal with the BEOL impact, we developed a timing-driven MIV planning flow and a post-route optimization flow to compensate for the reduced routing layers and increased resistance of tungsten interconnect. Experiments along with 7nm bulk FinFET from a foundry-grade

PDK demonstrated that our design solution allows only 3% performance degradation in G-M3D ICs under the worst FEOL/BEOL degradation scenario.

8.1.4 Compact-2D: A Physical Design Methodology to Build Commercial-Quality F2F 3D ICs

To maximize the utilization of 3D interconnect and the power-performance-area benefit of F2F-bonded 3D ICs, in this research, we proposed a full-chip RTL-to-GDSII physical design solution named Compact-2D (C2D) that offers a commercial-quality F2F-bonded 3D IC physical layout.

To sum up the strengths of our C2D flow, firstly C2D does not shrink the standard cell and interconnect geometries, so we can utilize the 2D P&R engines for the current technology node. Secondly C2D offers strong post-tier-partitioning optimization that enables timing, power, and F2F location co-optimization further. This makes C2D flow more favorable and adaptable in the advanced technology node. On the other hand, C2D requires the accurate parasitic database of the full 3D metal stack for the decent post-tier-partitioning optimization, which is challenging due to the limited support from tools and commercial PDKs for 2D ICs. To make C2D more powerful, the impact of multiple active layers on the interconnect parasitics, and the detailed die-to-die, die-to-F2F, and F2F-to-F2F couplings in addition to a simple F2F via parasitic model needs to be accounted. Also, along with placement row splitting, supporting full DRV fixing on the macro pins outside the boundaries will make post-tier-partitioning optimization more precise, and eventually remove the sign-off DRV fixing in the incremental routing stage.

With our extensive experiments and analysis, we evaluated the impact of those ideas in the final F2F design results, and showed that using 28nm process design kit, F2F-bonded 3D ICs implemented by our C2D flow offers a maximum 26.8% of total power reduction with a maximum 15.6% silicon area savings compared to the 2D IC designs at iso-performance.

8.1.5 Design and Architectural Co-optimization of M3D LSM Neuromorphic Processor

In this work, we implemented M3D IC design for an LSM-based neuromorphic processor and devised various design and architectural co-optimizations to minimize the energy consumption in the speech recognition. We presented the impact of shared memory architecture and the impact of the synaptic model complexity on the individual neuron and full-chip design. We measured the energy dissipation for speech recognition application with TI46 corpus spoken English speech samples, and achieved up to 70.0% reduction in the power-performance-area-accuracy overhead.

8.1.6 Area-efficient and Low-power Gate-level F2F 3D LSM Design

In this work, we studied the impact of the reservoir size and the connectivity density on the classification accuracy and their area-power overhead of the liquid state machine (LSM) processor designs. We showed that the 135 reservoir neurons, which is the maximum size of the reservoir used in this work, significantly improves the learning performance of a representative non-trivial real-world benchmark of speech recognition by 4.4% compared with the design with 72 reservoir neurons. However, we observed that this accuracy improvement is not for free and requires 78% more form factor, and 96.4% power consumption. Regarding the reservoir connectivity, we presented that denser reservoir improves the accuracy by a maximum 0.8% without increasing the network size, but this results in 28% more form factor, and 38% more power consumption in average.

Lastly, we explored the F2F-bonded 3D integration benefits on the LSM processor design using the state-of-the-art physical design flow name Compact-2D, and thoroughly analyzed the area-power benefits in the two-tier F2F-bonded 3D LSM processor design. We observed that F2F-bonded 3D integration has significant benefits on the form factor savings, and additional power savings while preserving the great classification accuracy. Using the design with 135 reservoir neurons, the F2F-bonded 3D LSM processor achieved 52% form factor savings, which is smaller than the 2D design with 72 liquid neurons by 13.8%,

and additional 3% power savings compared to its 2D counterpart. This work suggests the power-area-accuracy tradeoffs on the reservoir optimization in the 2D and F2F-bonded 3D LSM processors targeting the next generation neuromorphic edge-computing devices in the Internet-of-Things Era.

8.2 Future Directions

This research has focused on developing CAD and design solutions to build high-quality two-tier 3D ICs. However, as the number of vertically stacked tiers has no limit, the power-performance-area benefits of 3D ICs can be improved when we build multi-tier 3D ICs. One approach is to generalize proposed Compact-2D flow to handle more than two tiers. We can adjust the scaling factors in interconnect RC scaling / placement contraction considering the number of tiers. Various multi-way balanced partitioning schemes can be applied to the tier partitioning for the given implementation and fabrication constraints as well.

Another future direction of this research is to investigate the impact of emerging non-volatile memory (NVM) technologies on the 3D neurmorphic processor design. In a next-generation neuromorphic system, memristor-based NVM memories, such as phase-change RAM (PCRAM), spin-transfer-torque magnetoresistive RAM (STT-MRAM), and resistive RAM (RRAM) are expected to be used as advanced synaptic devices due to their efficient in-memory processing capability, and low energy consumption. Recently, many studies on NVM technologies present that the vertical integration of NVM significantly improves the memory density while being compatible to the traditional CMOS process. When NVM technologies are adopted to the multi-tier 3D ICs, this will bring us enormous synergy revolutionizing our lives.

REFERENCES

[1] G. E. Moore, "Cramming more components onto integrated circuits," *Proceedings of the IEEE*, vol. 86, no. 1, pp. 82–85, 1998.

[2] M. T. Bohr and I. A. Young, "Cmos scaling trends and beyond," *IEEE Micro*, vol. 37, no. 6, pp. 20–29, 2017.

[3] *International Roadmap for Devices and Systems 2017*, [Online; accessed 11-July-2018].

[4] Wikipedia contributors, *Transistor count — Wikipedia, the free encyclopedia*, [Online; accessed 10-July-2018].

[5] R. han Kim *et al.*, "Imec n7, n5 and beyond: dtco, stco and euv insertion strategy to maintain affordable scaling trend," vol. 10588, 2018, pp. 10588 –10588 –10.

[6] A. P. Jacob *et al.*, "Scaling challenges for advanced cmos devices," *International Journal of High Speed Electronics and Systems*, vol. 26, no. 01n02, p. 1 740 001, 2017.

[7] D. Yakimets *et al.*, "Vertical gaafets for the ultimate cmos scaling," *IEEE Transactions on Electron Devices*, vol. 62, no. 5, pp. 1433–1439, 2015.

[8] P. Raghavan *et al.*, "5nm: has the time for a device change come?" In *2016 17th International Symposium on Quality Electronic Design (ISQED)*, 2016, pp. 275–277.

[9] H. Mertens *et al.*, "Gate-all-around mosfets based on vertically stacked horizontal si nanowires in a replacement metal gate process on bulk si substrates," in *2016 IEEE Symposium on VLSI Technology*, 2016, pp. 1–2.

[10] E. Beyne, "The 3-d interconnect technology landscape," *IEEE Design Test*, vol. 33, no. 3, pp. 8–20, 2016.

[11] C. S. Tan, R. J. Gutmann, and L. R. Reif, *Wafer level 3-D ICs process technology*. Springer Science & Business Media, 2009.

[12] T. Suga *et al.*, "Direct cu to cu bonding and other alternative bonding techniques in 3d packaging," *Springer*, pp. 129–155, 2017.

[13] P. R. Morrow *et al.*, "Three-dimensional wafer stacking via cu-cu bonding integrated with 65-nm strained-si/low-k cmos technology," *IEEE Electron Device Letters*, vol. 27, no. 5, pp. 335–337, 2006.

[14] T. N. Theis and H. S. P. Wong, "The end of moore's law: a new beginning for information technology," *Computing in Science Engineering*, vol. 19, no. 2, pp. 41–50, 2017.

[15] P Batude *et al.*, "3-d sequential integration: a key enabling technology for heterogeneous co-integration of new function with cmos," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2012.

[16] S. W. Kim *et al.*, "Ultra-fine pitch 3d integration using face-to-face hybrid wafer bonding combined with a via-middle through-silicon-via process," in *Electronic Components and Technology Conference*, 2016, pp. 1179–1185.

[17] E. Beyne, "The 3-d interconnect technology landscape," *IEEE Design Test*, vol. 33, no. 3, pp. 8–20, 2016.

[18] D. K. Nayak, S. Banna, S. K. Samal, and S. K. Lim, "Power, performance, and cost comparisons of monolithic 3d ics and tsv-based 3d ics," in *SOI-3D-Subthreshold Microelectronics Technology Unified Conference*, 2015, pp. 1–2.

[19] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[20] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[22] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*, ACM, 2008, pp. 160–167.

[23] D. Verstraeten, B. Schrauwen, D. Stroobandt, and J. Van Campenhout, "Isolated word recognition with the liquid state machine: a case study," *Information Processing Letters*, vol. 95, no. 6, pp. 521–528, 2005.

[24] A. Ghani, T. M. McGinnity, L. P. Maguire, and J. Harkin, "Neuro-inspired speech recognition with recurrent spiking neurons," in *Artificial Neural Networks-ICANN 2008*, Springer, 2008, pp. 513–522.

[25] Y. Zhang, P. Li, Y. Jin, and Y. Choe, "A digital liquid state machine with biologically inspired learning and its application to speech recognition," *IEEE transactions on neural networks and learning systems*, vol. 26, no. 11, pp. 2635–2649, 2015.

[26] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Design and cad methodologies for low power gate-level monolithic 3d ics," in *Proc. Int. Symp. on Low Power Electronics and Design*, 2014, pp. 171–176.

[27] Y.-J. Lee, D. Limbrick, and S. K. Lim, "Power benefit study for ultra-high density transistor-level monolithic 3d ics," in *Proc. ACM Design Automation Conf.*, 2013, pp. 1–10.

[28] J. Shi *et al.*, "On the design of ultra-high density 14nm finfet based transistor-level monolithic 3d ics," in *IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 2016, pp. 449–454.

[29] B. W. Ku *et al.*, "How much cost reduction justifies the adoption of monolithic 3d ics at 7nm node?" In *Proceedings of the 35th International Conference on Computer-Aided Design*, ACM, 2016, p. 87.

[30] C. Fenouillet-Beranger *et al.*, "Fdsoi bottom mosfets stability versus top transistor thermal budget featuring 3d monolithic integration," *Solid-State Electronics*, vol. 113, pp. 2 –8, 2015.

[31] S. Sinha, G. Yeric, V. Chandra, B. Cline, and Y. Cao, "Exploring sub-20nm finfet design with predictive technology models," in *Proc. ACM Design Automation Conf.*, 2012, pp. 283–288.

[32] *International technology roadmap for semiconductors 2013*, 2013.

[33] K. Bhanushali and W. R. Davis, "Freepdk15: an open-source predictive process design kit for 15nm finfet technology," in *Proceedings of the 2015 Symposium on International Symposium on Physical Design*, ser. Proc. Int. Symp. on Physical Design, 2015, pp. 165–170.

[34] L. Liebmann, J. Zeng, X. Zhu, L. Yuan, G. Bouche, and J. Kye, "Overcoming scaling barriers through design technology cooptimization," in *IEEE Int. Symposium on VLSI Technology, Systems, and Applications*, 2016, pp. 1–2.

[35] K. Miyaguchi *et al.*, "Modeling finfet metal gate stack resistance for 14nm node and beyond," in *International Conference on IC Design Technology (ICICDT)*, 2015, pp. 1–4.

[36] Y. Sasaki *et al.*, "Novel junction design for nmos si bulk-finfets with extension doping by peald phosphorus doped silicate glass," in *Proc. IEEE Int. Electron Devices Meeting*, 2015, pp. 21.8.1–21.8.4.

[37] Y. S. Yu, S. Panth, and S. K. Lim, "Electrical coupling of monolithic 3-d inverters," *IEEE Transactions on Electron Devices*, vol. 63, pp. 3346–3349, 2016.

[38] M. Martins *et al.*, "Open cell library in 15nm freepdk technology," in *Proc. Int. Symp. on Physical Design*, 2015, pp. 171–178.

[39] R. Aitken *et al.*, "Physical design and finfets," in *Proc. Int. Symp. on Physical Design*, 2014, pp. 65–68.

[40] B. Chava *et al.*, "Standard cell design in n7: euv vs. immersion," in *Proc. SPIE*, 2016, 94270E–94270E–9.

[41] http://www.opencores.org/.

[42] K. Chang, K. Acharya, S. Sinha, B. Cline, G. Yeric, and S. K. Lim, "Power benefit study of monolithic 3d ic at the 7nm technology node," in *Proc. Int. Symp. on Low Power Electronics and Design*, 2015, pp. 201–206.

[43] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Placement-driven partitioning for congestion mitigation in monolithic 3d ic designs," *IEEE Trans. on Computer-Aided Design of Int. Circuits and Systems*, vol. 34, no. 4, pp. 540–553, 2015.

[44] A. Mallik *et al.*, "Maintaining moore's law: enabling cost-friendly dimensional scaling," vol. 9422, 2015, 94221N–94221N–12.

[45] X. Dong, J. Zhao, and Y. Xie, "Fabrication cost analysis and cost-aware design space exploration for 3-d ics," vol. 29, no. 12, pp. 1959–1972, 2010.

[46] Q. Zou, J. Xie, and Y. Xie, "Cost-driven 3d design optimization with metal layer reduction technique," in *Proc. Int. Symp. on Quality Electronic Design*, 2013, pp. 294–299.

[47] A. Mallik *et al.*, "The need for euv lithography at advanced technology for sustainable wafer cost," vol. 8679, 2013, 86792Y–86792Y–10.

[48] ——, "The economic impact of euv lithography on critical process modules," in *Proc. SPIE*, vol. 9048, 2014, 90481R–90481R–12.

[49] P. Batude *et al.*, "3d sequential integration opportunities and technology optimization," in *Proc. IEEE Int. Interconnect Technology Conference*, 2014, pp. 373–376.

[50] F. Luce *et al.*, "Methodology for thermal budget reduction of sper down to 450c for 3d sequential integration," *Nuclear Instruments and Methods in Physics Research*, vol. 370, pp. 14 –18, 2016.

[51] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Power-performance study of block-level monolithic 3d-ics considering inter-tier performance variations," in *Proc. ACM Design Automation Conf.*, 2014, pp. 1–6.

[52] D. H. Kim *et al.*, "Design and analysis of 3d-maps (3d massively parallel processor with stacked memory)," *IEEE Transactions on Computers*, vol. 64, no. 1, pp. 112–125, 2015.

[53] C. S. Premachandran *et al.*, "A novel, wafer-level stacking method for low-chip yield and non-uniform, chip-size wafers for mems and 3d sip applications," in *Electronic Components and Technology Conference*, 2008, pp. 314–318.

[54] B. W. Ku *et al.*, "Physical design solutions to tackle feol/beol degradation in gate-level monolithic 3d ics," in *Proc. Int. Symp. on Low Power Electronics and Design*, 2016, pp. 76–81.

[55] K. Chang *et al.*, "Cascade2d: a design-aware partitioning approach to monolithic 3d ic with 2d commercial tools," in *Proc. IEEE Int. Conf. on Computer-Aided Design*, 2016, 130:1–130:8.

[56] http://www.opensparc.net/.

[57] M. LukošEvičIus and H. Jaeger, "Reservoir computing approaches to recurrent neural network training," *Computer Science Review*, vol. 3, no. 3, pp. 127–149, 2009.

[58] Q. Wang, Y. Li, and P. Li, "Liquid state machine based pattern recognition on FPGA with firing-activity dependent power gating and approximate computing," in *Internatioal Symposium of Circuits and Systems (ISCAS), 2016 IEEE*, IEEE, 2016, pp. 361–364.

[59] Y. Jin, Y. Liu, and P. Li, "Sso-lsm: a sparse and self-organizing architecture for liquid state machine based neural processors," in *Nanoscale Architectures (NANOARCH), 2016 IEEE/ACM International Symposium on*, IEEE, 2016, pp. 55–60.

[60] Y. Liu, Y. Jin, and P. Li, "Exploring sparsity of firing activities and clock gating for energy-efficient recurrent spiking neural processors," in *Low Power Electronics and Design (ISLPED, 2017 IEEE/ACM International Symposium on*, IEEE, 2017, pp. 1–6.

[61] TI46, *The TI46 speech corpus.* `http://catalog.ldc.upenn.edu/LDC93S9`.

[62] R. F. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82.*, IEEE, vol. 7, 1982, pp. 1282–1285.

[63] B. Schrauwen and J. Van Campenhout, "BSA, a fast and accurate spike train encoding scheme," in *Proceedings of the International Joint Conference on Neural Networks*, IEEE Piscataway, NJ, vol. 4, 2003, pp. 2825–2830.

[64] E. Beyne, "The 3-d interconnect technology landscape," *IEEE Design & Test*, vol. 33, no. 3, pp. 8–20, 2016.

[65] B. W. Ku, K. Chang, and S. K. Lim, "Compact-2d: a physical design methodology to build commercial-quality face-to-face-bonded 3d ics," in *Proc. Int. Symp. on Physical Design*, 2018.

[66] W. Maass, T. Natschläger, and H. Markram, "Real-time computing without stable states: A new framework for neural computation based on perturbations," *Neural computation*, vol. 14, no. 11, pp. 2531–2560, 2002.

PUBLICATIONS

This dissertation is based on and/or related to the works and results presented in the following publications in print:

[1] Kartik Acharya, Kyungwook Chang, **Bon Woong Ku**, Shreepad Panth, Saurabh Sinha, Brian Cline, Greg Yeric, and Sung Kyu Lim, "Monolithic 3D IC Design: Power, Performance, and Area Impact at 7nm," in *IEEE International Symposium on Quality Electronic Design*, 2016, pp. 41-48.

[2] **Bon Woong Ku**, Peter Debacker, Dragomir Milojevic, Praveen Raghavan, Diederik Verkest, Aaron Thean and Sung Kyu Lim, "Physical Design Solutions to Tackle FEOL/BEOL Degradation in Gate-level Monolithic 3D ICs," in *ACM International Symposium on Low Power Electronics and Design*, 2016, pp. 76-81.

[3] **Bon Woong Ku**, Dragomir Milojevic, Peter Debacker, Praveen Raghavan, and Sung Kyu Lim, "How Much Cost Reduction Justifies the Adoption of Monolithic 3D ICs at 7nm Node?," in *ACM International Conference on Computer Aided Design*, 2016, pp. 87-92.

[4] **Bon Woong Ku**, Taigon Song, Arthur Nieuwoudt, and Sung Kyu Lim, "Transistor-Level Monolithic 3D Standard Cell Layout Optimization for Full-Chip Static Power Integrity," in *IEEE International Symposium on Low Power Electronics and Design*, 2017, pp. 1-6.

[5] A. Mallik, A. Vandooren, L. Witters, A. Walke, J. Franco, Y. Sherazi, D. Yakimets, MG Bardon, B. Parvais, P. Debacker, **Bon Woong Ku**, S. K. Lim, A. Mocuta, D. Mocuta, J. Ryckaert, N. Collaert, P. Raghavan, "The Impact of Sequential-3D Integration on Semiconductor Scaling Roadmap," in *IEEE International Electron Devices*

Meeting, 2017, pp. 32.1.1-31.1.4.

[6] Kyungwook Chang, **Bon Woong Ku**, Saurabh Sinha, and Sung Kyu Lim, "Full-chip Monolithic 3D IC Design and Power Performance Analysis with ASAP7 Library: (Invitied Paper)," in *IEEE International Conference on Computer-Aided Design*, 2017, pp. 1005-1010.

[7] **Bon Woong Ku**, Kyungwook Chang, and Sung Kyu Lim, "Compact-2D: A Physical Design Methodology to Build Commercial-Quality Face-to-Face-Bonded 3D ICs," in *ACM International Symposium on Physical Design*, 2018, pp. 90-97.

[8] **Bon Woong Ku**, Yu Liu, Yingyezhe Jin, Sandeep Samal, Peng Li, and Sung Kyu Lim, "Design and Architectural Co-optimization of Monolithic 3D Liquid State Machine-based Neuromorphic Processor," in *IEEE Design Automation Conference*, 2018, pp. 1-6.

[9] **Bon Woong Ku**, Yu Liu, Yingyezhe Jin, Peng Li, and Sung Kyu Lim, "Area-efficient Low-power Face-to-Face-bonded 3D Liquid State Machine Design in the Internet-of-Things Era," in *IEEE/ACM International Conference on Computer Aided Design*, 2018, pp. 1-6.

[10] **Bon Woong Ku**, Kyungwook Chang, and Sung Kyu Lim, "Compact-2D: A Physical Design Methodology to Build Two-Tier Gate-level 3D ICs," *IEEE Transaction on Computer-Aided Design of Integrated Circuits and Systems*, to appear.

In addition, the author has completed works unrelated to this dissertation presented in the following publications in print:

[1] Yarui Peng, **Bon Woong Ku**, Younsik Park, Kwang-Il Park, Seong-Jin Jang, Joo Sun Choi, and Sung Kyu Lim, "Design, Packaging, and Architectural Policy Co-Optimization for DC Power Integrity in 3D DRAM," in *ACM Design Automation Conference*, 2015, pp. 91-96.

[2] Catherine Schuman, Raphael Pooser, Tiffany Mintz, Md Musabbir Adnan, Garrett Rose, **Bon Woong Ku** and Sung Kyu Lim, "Simulating and Estimating the Behavior of a Neuromorphic Co-Processor," in *International Workshop on Post-Moore's Era Supercomputing*, Denver, CO, November 2017, pp. 8-14.

[3] Austin Wyer, Md Musabbir Adnan, **Bon Woong Ku**, Sung Kyu Lim, Catherine D. Schuman, Raphael C. Pooser and Garrett S. Rose, "Evaluating Online-Learning in Memristive Neuromorphic Circuits," in *Neuromorphic Computing Workshop: Architectures, Models, and Applications*, Oak Ridge, TN, July 2017.

[4] **Bon Woong Ku**, MD Musabbir Adnan, Catherine D. Schumann, Tiffany Mintz, Raphael Pooser, Garrett S. Rose, and Sung Kyu Lim, "Stochastic Digital Spike-timing-dependent Plasticity Implementation for Memristive Neuromorphic System: (Extended Abstract)," in *ACM International Conference on Neuromorphic Systems*, Oak Ridge, TN, 2018.

[5] Md Musabbir Adnan, Sagarvarma Sayyaparaju, Catherine D. Schuman, **Bon Woong Ku**, Sung Kyu Lim, and Garrett Rose, "A Twin Memristor Synapse for Spike Timing Dependent Learning in Neuromorphic Systems," in *IEEE International System-on-Chip Conference*, 2018, pp. 37-42.

[6] **Bon Woong Ku**, Catherine D. Schuman, Tiffany M. Mintz, Raphael Pooser, Md Mussabir Adnan, Kathleen E. Hamilton, Garrett S. Rose, and Sung Kyu Lim, "Unsupervised Digit Recognition Using Single-Spike Temporal Coding In A Neuromemristive Competitive Learning System," *IEEE Transactions on Neural Networks and Learning Systems*, to appear.

VITA

Bon Woong Ku was born in Suwon, South Korea in 1990. He received the B.S. degree from Seoul National University, Seoul, South Korea in 2014, and the M.S. degree from Georgia Institute of Technology, Atlanta, GA, USA, in 2017, where he is currently a Ph. D. candidate. From 2014 up to the present, he has been a graduate research assistant in Georgia Tech Computer Aided Design (GTCAD) laboratory led by Dr. Sung Kyu Lim. His primary research interests include emerging device modeling and aspects of physical design and CAD solution for next generation 3D ICs in the advanced technology and its neuromorphic computing applications.